AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

D6.2

Report on Policy for Content Moderation

Project Title	AI4Media - A European Excellence Centre for Media, Society and Democracy
Contract No.	951911
Instrument	Research and Innovation Action
Thematic Priority	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) / ICT-48-2020 - Towards a vibrant European network of AI excellence centres
Start of Project	1 September 2020
Duration	48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Deliverable title	Report on Policy for Content Moderation
Deliverable number	D6.2
Deliverable version	1.0
Previous version(s)	-
Contractual date of delivery	28 February 2023
Actual date of delivery	22 March 2023
Deliverable filename	D6.2_Report_on_Policy_for _Content_Moderation.docx
Nature of deliverable	Report
Dissemination level	Public
Number of pages	130
Work Package	WP6
Task(s)	T6.1
Partner responsible	KUL
Author(s)	Noémie Krack (KUL), Lidia Dutkiewicz (KUL), Emine Ozge Yildirim (KUL)
Editor	Noémie Krack (KUL)
EC Project Officer	Evangelia Markidou

Abstract	Deliverable 6.2 "Report on Policy for Content Moderation" presents an overview of the EU policy initiatives on content moderation as well alternative approaches to content moderation by online platforms and civil society. It assesses the challenges and advantages of these instruments and diverging approaches to outline policy recommendations for the future of content moderation in the EU landscape. More specifically, D6.2 provides an introduction to content moderation, including algorithmic content moderation and its challenges to fundamental rights such as freedom of expression, as well as an analysis of the legal landscape composed of hard law (lex generalis and lex specialis) and other types of regulatory instruments. It investigates the criticisms addressed to each of these instruments and recommendations for the future. It also analyses self-regulatory initiatives as alternative approaches, such as end-user moderation and self-moderation through bodies and new models. Moreover, it reflects on the AI4Media workshop on AI and content moderation held with media practitioners. Finally, based on the results of the previous analysis, it provides a set of policy recommendations for content moderation.
Keywords	Content moderation, Law, Media, AI, self-regulation, hard regulation, fundamental rights, policy recommendations, algorithmic content moderation.

www.ai4media.eu

info@ai4media.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Copyright

Contributors

NAME	ORGANISATION
NOEMIE KRACK	KUL
LIDIA DUTKIEWICZ	KUL
EMINE OZGE YILDIRIM	KUL

Peer Reviews

NAME	ORGANISATION
SVEN BECKER	FHG-IAIS
GEORGI KOSTADINOV	IMG
HRISTO GEORGIEV	IMG
ALEKSANDRA KUCZERAWY	KUL
FILARETI TSALAKANIDOU	CERTH

Revision History

VERSION	DATE	REVIEWER	MODIFICATIONS
0.1	01/12/2022	Noémie Krack, Lidia Dutkiewicz, Emine Ozge Yildirim	1st table of content
0.2	05/02/2023	Noémie Krack, Lidia Dutkiewicz, Emine Ozge Yildirim	1st draft
0.3	31/03/2023	Noémie Krack, Lidia Dutkiewicz, Emine Ozge Yildirim	2nd draft following the AI4Media workshop on AI and content moderation.
0.4	08/03/2023	Noémie Krack, Lidia Dutkiewicz, Emine Ozge Yildirim	Version ready to be sent to the internal reviewers.
0.5	17/03/2023	Sven Becker, Georgi Kostadinov, Hristo Georgiev, Aleksandra Kuczerawy, Filareti Tsalakanidou	Version with reviewers comments.

1.0	**21/03/2023**	**Noémie Krack, Lidia Dutkiewicz, Emine Ozge Yildirim**	**Final version after internal review comments were addressed.**

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Table of Abbreviations and Acronyms

Abbreviation	Meaning
3D	Three-dimensional
AI	Artificial Intelligence
API	Application Programming Interface
Art.	Article
AVMSD	Audiovisual Media Service Directive
CCDH	Center for Countering Digital Hate
CC BY-SA	Creative Commons Attribution-ShareAlike
CDSM	Copyright in the Digital Single Market Directive
CJEU	Court of Justice of the European Union
Council	Council of the European Union
CSAED	The Child Sexual Abuse and Exploitation Directive
CSAM	Child Sexual Abuse Material
D.	Deliverable
Dir.	Directive
DMA	Digital Market Act
DMCA	US Digital Millennium Copyright Act
DPPA	United Nations' Department of Political and Peacebuilding Affairs
DSA	Digital Services Act
EC	European Commission
ECD	e-commerce Directive
ECFR	European Union Charter on Fundamental Rights
ECHR	European Convention on Human Rights
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EDRi	European Digital Rights
EECC	European Electronic Communication Code
EP	European Parliament
EU	European Union
EUIF	European Union Internet Forum
FOB	Facebook Oversight Board
GDPR	General Data Protection Regulation
GIFCT	Global Internet Forum to Counter Terrorism
HSP	Hosting Service Provider
ICC	International Criminal Court
IP	Intellectual Property
IRU	EU Internet Referral Unit
IT	Information Technology

Abbreviation	Meaning
KPI	Key Performance Indicator
LGBTQ+	Lesbian, Gay, Bisexual, Transgender, Queer and others
MEP	Member of the European Parliament
MS	Member State
NASA	National Aeronautics and Space Administration
NGO	Non-Governmental Organisation
OB	Oversight Board
OCSSP	Online Content-Sharing Service Provider
OJ	Official Journal
OSP	Online Services Provider
PR	Public Relations
PTSD	Post-traumatic Stress Disorder
P2P	Peer to Peer
Rec.	Recital
Reg.	Regulation
SMC	Social Media Council
T&C	Terms & Conditions
TERREG	Regulation on preventing the dissemination of terrorist content online
TEU	Treaty on the European Union
TFEU	Treaty on the Functioning of the European Union
TOS	Terms of Services
UK	United Kingdom
UN	United Nations
UGC	User Generated Content
UGV	User Generated Video
URL	Uniform Resource Locator
US	United States
VLOP	Very Large Online Platform
VLOSE	Very Large Online Search Engine
VR	Virtual Reality
VSP	Video-Sharing Platform
WMF	Wikimedia Foundation

Index of Contents

Index of Tables

Index of Figures

1 Executive Summary

Deliverable D6.2 *"Report for Policy on Content Moderation"* provides an overview of the EU policy initiatives on content moderation and the future trends and alternative approaches to content moderation by online platforms. By doing so, D6.2 aims to assess the challenges and advantages of these instruments and diverging approaches and to outline a set of policy recommendations for the future of content moderation in the EU landscape.

Section 3 explains the primary policy and legal instruments in the content moderation landscape in a thematic manner. More specifically, Section 3.1 provides introductory remarks on the concept of content moderation. This includes what content moderation is, how algorithmic content moderation is utilised and what are its technical limitations (including lack of context, quality, diversity, and inclusivity), and what are the relevant socio-political challenges that such automation poses. Then, it further elaborates on algorithmic content moderation's challenges such as the right to freedom of expression and other fundamental rights.

Section 3.2 analyses the overarching instruments in the EU legal landscape concerning content moderation along two dimensions. First, the horizontal rules, which apply to all types of content (lex generalis), are outlined in detail. This includes the e-Commerce Directive, the newly adopted Digital Services Act (DSA) applicable to online platforms in general, and the Audio-Visual Media Services Directive (AVMSD) that imposes obligations on video-sharing platforms. Next, it focuses on the vertical rules, which apply to specific types of content deserving specific attention, rules, and processes (lex specialis). Such specific content includes terrorist content, child sexual abuse material (CSAM), copyright infringing content, racist and xenophobic content, disinformation, and hate speech. Not all this content could be deemed illegal, as there are different sensitivity degrees concerning each content. Therefore, lex specialis regulations consist of specific rules applicable to different types of content, and they are also sometimes complemented by soft law instruments. The following lex specialis instruments are discussed in detail: the EU Regulation against terrorist content online, the Directive on Copyright in the Digital Single Market, the Child Sexual Abuse and Exploitation Directive, the Interim CSAM Regulation, and the proposal for a new regulation combatting CSAM. Moreover, soft-law instruments such as the EU Code of Conduct on countering illegal hate speech online, the Code of Practice on disinformation, and the strengthened Code of Practice on Disinformation are further elaborated on.

Apart from regulatory instruments, some platforms also follow their own approaches to content moderation. These alternative approaches, which could also be called self-regulatory initiatives, include end-user moderation, community-led moderation, and accountability mechanisms established by some platforms. To this end, section 4 provides an analysis of these diverging approaches in the following manner:

Section 4.1.1 explains how voluntary editing and moderation are implemented on self-moderated community platforms such as Wikipedia and Discord and whether the DSA would apply to their self-regulatory mechanisms. Section 4.1.2 outlines the content moderation practices in the Fediverse, which refers collectively to the protocols, servers, and applications that enable decentralised social media. The section uses the Mastodon project as a case study to expand on this under-researched content moderation practice. Furthermore, Section 4.1.3 analyses the content moderation in the metaverse, which could be described as an immersive 3D world, also from the point of view of the challenges Virtual Reality and Augmented Reality technologies could pose to content moderation in the near future. Additionally, Section 4.2 outlines the advantages and challenges of self-regulatory accountability mechanisms such as the Facebook Oversight Board and the civil society-proposed Social Media Councils.

In order to bridge the gap between academia and the industry, AI4Media has been seeking opportunities for further networking and discussion of critical questions. With that aim, a workshop on AI and Content Moderation was organised by two consortium partners – KUL and UvA – inviting distinguished academics, companies developing tools for content moderation, journalism, and newspaper companies, a representative of a very large online platform, and a consultant from an intergovernmental organisation as participants. Therefore, in Section 5, the main takeaways and the results of the workshop are explained extensively.

After identifying the main challenges concerning content moderation in Sections 3, 4, and 5, Section 6 offers a set of policy recommendations (high-level as well as problem-specific), mostly addressed to the EU policymakers, also targeting European policymakers at large. The main horizontal and high-level recommendations (Section 6.1) are summarised below:

- Utilise a tailored combination of regulatory instruments, technologies, and content moderation approaches to fit specific content and context needs.
- Increase communication, awareness, and compliance support for the complex EU regulatory landscape targeting all stakeholders.
- Investigate which technologies and approaches work best for different types of content and contexts, taking into account geographical, language, and diverse communities.
- Tailor the use of chosen technologies and approaches based on the type of content being moderated, such as text, image, live stream, and the like.
- Ensure transparency and safeguards in content moderation sub-contracting and human moderator working conditions.
- Enforce new tech legislation impacting content moderation, such as the Digital Services Act and Digital Markets Act provisions for transparency and access, to improve accountability and enable a better understanding of content moderation mechanisms.

After these overarching and high-level recommendations, Section 6.2 provides recommendations concerning specific content such as:

- Terrorist content,
- Copyright-protected content,
- Child sexual abuse material,

- Hate speech, and
- Disinformation.

Finally, Section 7 concludes the deliverable. Despite policymakers' attempts to regulate the growing power of big tech platforms and their impact on freedom of expression online, content, and the use of technology, achieving coherence among various legal instruments and striking a delicate balance between competing interests and fundamental rights presents challenges. There are also different approaches taken by private actors to address content moderation challenges. However, there is no single way to address the multi-complexity of content moderation. Therefore, an encompassing approach that takes into account the specificities of different types of content, actors, and services is necessary. To ensure the sustainability of the online realm, it is also essential to enhance transparency in vital areas such as institutional transparency, infrastructure, and the labour market. Additionally, it is necessary to scrutinize less represented forms of illegal/harmful content to promote transparency further. To that end, more research is needed to ensure that content moderation initiatives, both by regulatory and private parties, are designed in balance with all values, rights, and interests at stake.

2 Introduction

One of the goals of Task 6.1 “*Policy recommendations for content moderation*” is to assess different aspects of future regulation of content moderation, revolving around the debate of whether regulatory instruments or self-regulation is more promising to ensure respect for fundamental rights without infringing upon the open public debate. The answer to this question is a patchwork, as both of these approaches have promising features, along with the substantial challenges and risks they pose.

The Internet surely changed the way of communication and broke down the traditional barriers to entry into the market, resulting in a massive boom of social media platforms and networks, ease of creating content, speedy dissemination of user-generated content, and (almost) untethered access to knowledge. The concept of ‘cheap speech’ initially looked so promising, as it had the potential to allow for a lively debate in the marketplace of ideas. However, the technological shift has moved the online sphere much further than the initial promise, perhaps to an unimagined land of real dangers and threats, requiring this space to be regulated to make it safer for users and tackle with power asymmetries and unlimited power platforms could possess. As a result, starting from the initial boom of social media platforms in the late 90s, the policymakers all around the world have been seeking to regulate the online sphere. In the EU, the regulatory efforts started with the e-Commerce Directive in 2000. The efforts have intensified in the last few years, as the possible threats of not moderating content online became more obvious nowadays. Additionally, with the objective of establishing a digital single market encompassing a cross-border phenomenon such as the Internet, the EU seeks to harmonise the legal landscape on content moderation while maintaining a consistent approach to regulating the online sphere.

There are several approaches envisaged for content moderation, including self-regulation, co-regulation, and hard regulation. Each approach has its own advantages and challenges, as content moderation is a complex subject at the crossroads where different fundamental rights meet, including freedom of expression, privacy and data protection, non-discrimination, freedom of thought, and the like. This is also due to the power imbalance and lack of legitimacy concerning private platforms’ attempts to regulate online public debate. It is crucial to acknowledge that online public debate in public forums plays a significant part in shaping public opinion. Users rely on these platforms to obtain information that helps them make informed decisions, whether as democratic citizens or consumers. When it comes to state authorities, they usually face a shortage of technical resources, as well as financial and human resources, to monitor content online. They also have a distinct agenda from commercial entities. Moreover, since it is undesirable for these authorities to act as the sole arbitrators of truth or engage in excessive surveillance and removal of online content that could suppress free expression, the concept of state-controlled content moderation is deemed less than ideal. Apart from the risk of suppressing free expression, there might also be other shortcomings that could appear with the intensified efforts of regulating in the EU. This also includes the issue of lagging and addressing risks once they already materialise, whereas it is crucial to act very precisely and in a sound way in light of the direct impact of interventions and non-interventions in this matter.

To the ends mentioned above, this deliverable aims to evaluate and analyse existing and upcoming regulations related to online content moderation. It also seeks to provide tailored policy recommendations that would contribute to the existing literature and influence future research on content sharing and the evolution of online content moderation. D6.2 also builds upon D6.1 "*First generation of Human- and Society-centered AI algorithms*", where KUL provided an initial analysis of the content moderation legal landscape (see Section 3 of D6.1).

The deliverable follows the structure below:

- **Section 3**, the Evolution of the EU Content Moderation Regulation, provides an analysis of the EU policy documents on content moderation.
    - Section 3.1, Content Moderation, explains what content moderation and algorithmic content moderation are, the challenges and limitations of automation in content moderation, and algorithmic content moderation challenges for freedom of expression and other fundamental rights.
    - Section 3.2, Content Moderation Landscape, maps the lex generalis and lex specialis policy and legal instruments in the EU, as well as the soft-law instruments, concerning content moderation:
        - Section 3.2.1, Horizontal Rules, outlines the lex generalis instruments in the EU, namely the e-Commerce Directive, the Digital Services Act, and the Audiovisual Media Service Directive.
        - Section 3.2.2, Vertical Rules Applicable to Illegal Content and Harmful Content, outlines the lex specialis and soft-law instruments on specific content such as terrorist content, copyright-protected content, child sexual abuse material, hate speech, and disinformation.
- **Section 4**, Alternative Approaches and Future Trends in Content Moderation, evaluates the divergent approaches by private actors and platforms themselves in content moderation:
    - Section 4.1, End-user Moderation or Community-led Moderation, outlines approaches in community-driven moderation in the following subsections:
        - Section 4.1.1, Self-moderated Communities, explains the self-regulatory and voluntary moderation community approach of platforms such as Wikipedia and Discord.
        - Section 4.1.2, Content Moderation in fediverse, analyses the decentralised nature of fediverse content moderation while using the Mastodon project as a case study.

        - Section 4.1.3, Content Moderation in the metaverse, focuses on the challenges and possible advantages of the metaverse in content moderation.
    - Section 4.2, Accountability Initiatives, explores alternative accountability initiatives proposed or conducted by online platforms or civil society:
        - Section 4.2.1, Facebook Oversight Board, gives an overview of the FOB's working mechanisms and provides a critical analysis of this approach.
        - Section 4.2.2, Social Media Councils, explains the social media councils' accountability model proposed by civil society and elaborates further on the advantages and challenges of this mechanism.
- **Section 5**, Workshop on AI and Content Moderation, provides an extensive overview of the workshop organised by KUL and UvA in February 2023, along with the main takeaways from the workshop and the recommendations and results incorporated.
- **Section 6**, Policy Recommendation on Content Moderation, outlines a set of policy recommendations, targeting the EU policymakers, commensurate to the challenges identified in previous sections, consisting of horizontal and high-level regulations, as well as more problem and content-specific recommendations.
- **Section 7**, Conclusions, ends with the final thoughts for future initiatives and research directions.

3 Evolution of the EU content moderation regulation

Considering the internet development, the invention and massive use of social media platforms and networks, the creation of user-generated content, and the mass of services, products, and content online, the question of whether to moderate content online has emerged quite rapidly. The benefits of uninterrupted access to the Internet are undeniable. Originally considered as an (almost) unrestricted place of freedom of speech, the zero restrictions environment would lead to a space filled with illegal, abusive speech that prevents regular users from participating in the debate and exchange of views. With time, the monitoring of the online environment by the Trust and Safety teams from tech companies and a legal and regulatory framework appeared necessary to make the online space a safer space for every online user. Thus, this section will dive into the concept of content moderation. It will show how AI is used to support and assist content moderation efforts and how the EU regulatory content moderation landscape has evolved up until today. The purpose of this section is to present the criticisms and propose suggestions for the future regarding the legislative and non-legislative texts governing the EU content moderation landscape.

3.1 Content moderation[1]

3.1.1 What is content moderation?

Content moderation definition

Internet intermediaries, typically, are private entities that provide commercial and technical infrastructure that allows information to be exchanged. Because of their enabling role and technical capabilities to affect, directly and indirectly, the behaviour and content of their users, they hold a powerful position and gather a considerable amount of online content, including user-generated content. These "internet information gatekeepers"[2] can eliminate access to a particular service, remove content, and amplify or downgrade information they choose to present.[3] In a broad sense, content moderation may therefore be understood as the "governance mechanisms that structure participation in a community to facilitate cooperation

[1] This section includes an updated version of the content provided in section 3 of deliverable D6.1 - "First-generation of Human- and Society-centered AI algorithms". Available here: https://www.ai4media.eu/reports/first-generation-of-human-and-society-centered-ai-algorithms-d6-1/.

[2] Emily B Laidlaw, 'A Framework for Identifying Internet Information Gatekeepers' (2010) 24 International Review of Law, Computers & Technology 263.

[3] Aleksandra Kuczerawy, 'Safeguards for Freedom of Expression in the Era of Online Gatekeeping' (20180914) 2017 Auteurs en Media 292.

and prevent abuse."[4] For the legal definition of content moderation in the Digital Services Act (DSA), see Section 3.2.1.3.

Content moderation occurs on many levels. It can take place before content is actually published on the website (*ex-ante* moderation) or after content is published (*ex-post* moderation). Moreover, moderators can passively assess content only after others flag the content to their attention, or they can proactively seek out published content. Additionally, content moderation decisions can be made either by automated (AI) means or manually by human content moderators. Often these two techniques go hand in hand. (See Section 3.1.2 for more information on automated content moderation). Recently, due to technological developments, content moderation has become a real market, as the immensity of user-generated content led to the creation of new businesses and jobs.

Content moderation value chain

To better understand content moderation dynamics, it is important to look at the content moderation value chain and the different actors involved in it. The first group of actors is **intermediary services**. Different types of intermediary service providers exist, such as caching, mere conduit, and hosting services. These different actors now own different kinds of obligations in light of the lex generalis and lex specialis of content moderation regulation (see Section 3.2). However, most obligations and responsibilities fall on the hosting services providers.

As their name indicates, **hosting services providers** host content on their services. Such content can also be user-generated content, including content itself, comments, sharing content, and re-posts, leading to many content moderation activities. Within their services, they will typically have teams handling the different components of content moderation. The **policy team** is responsible for following and implementing the new legislation or non-binding initiatives to which the company committed. They are also in charge of developing the company's own policies for handling different types of content. The policy team will design the terms of use and set up the rules determining what content will remain online and what will be taken down. The **operations team**'s activities involve setting up staffing, processes for taking down content, complaints handling, and reuploading content. They are executing the decisions and policies from the Policy team. People with a technical background, such as **developers, computer scientists, and engineers**, will be responsible for building the tools and infrastructures used at different times and by/for different actors of the content moderation value chain. Combined, these three function groups can be referred to as the **Trust and Safety** teams.[5]

[4] James Grimmelmann, 'The Virtues of Moderation' (LawArXiv 2017) preprint <https://osf.io/qwxf5> accessed 1 December 2021.

[5] 'What Is Content Moderation?' (*Trust and Safety Professional Association*) <https://www.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/what-is-content-moderation/> accessed 14 February 2023.

There are different sizes of intermediary services (large, middle, and small), from Facebook to local newspaper websites. Each category has access to different economic, human, and infrastructure resources and faces different content moderation challenges. Some have reported that "all major platforms have their own content moderation systems, often sub-contracted to specialised companies, to identify and take down online content"[6] deemed in conflict with their terms of services (ToS).

The second group consists of the **sub-contractors**. Companies can rely on third-party vendor models to perform content moderation activities. Relying on outsourced content moderation has been found to own quite some benefits for the company of origin. The outsourcing model allows the transfer of content moderation burden on experienced companies having their core business and focus on this specifically. This enables relying on the latest technologies, skilled people, training, and processes already in place, which can prove to be more time- and cost-efficient. These specialised content moderation companies often rely on both AI tools and human expertise. In practice, however, the outsourcing of content moderation by very large online platforms has been subject to some controversies (see below 'Opaqueness of the content moderation infrastructure'). Another trend includes acquiring content moderation companies through online platforms to strengthen in-house expertise.[7]

The third group consists of **content moderators** who are doing the human review part of the content moderation process.[8] They are enforcing platforms' terms of service or content guidelines. Some authors report that there are around 100.000 content moderators.[9] It is not always easy to find relevant information on platform's service about their moderators. In addition, opaque procedures, few or non-monitoring, and non-disclosure agreements part of the moderator employment agreements are preventing from getting a better understanding of the systems in place.[10] However, the traumatic impact of the content moderator work has been

[6] Sarah T Roberts, *Behind the Screen* (Yale University Press 2021) <https://yalebooks.yale.edu/9780300261479/behind-the-screen> accessed 10 February 2023.

[7] Spotify, 'Spotify Continues to Ramp Up Platform Safety Efforts with Acquisition of Kinzen' (*Spotify*, 5 October 2022) <https://newsroom.spotify.com/2022-10-05/spotify-continues-to-ramp-up-platform-safety-efforts-with-acquisition-of-kinzen/> accessed 22 February 2023.

[8] We will not go in depth on this subject, though there would be much to tell on this matter. The following article investigates the topic. Miriah Steiger and others, 'The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3411764.3445092> accessed 20 January 2023.

[9] Steiger and others (n 8).

[10] Andrew Arsht and Etcovitch, 'The Human Cost of Online Content Moderation' (*Harvard Journal of Law & Technology*, 2 March 2018) <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation> accessed 17 February 2023.

underlined by several investigations. In 2018, the documentary "The Cleaners"[11] depicts the work experience of a content moderator and the long-term psychological trauma associated with the job. Sarah Roberts also demonstrated how the factory-like routine of content moderation has led to burnout and mental distress such as post-traumatic stress disorder (PTSD).[12] It can also lead to a normalisation of violence and undermining people's trauma.

The documentary also shows how content moderators take down content for grounds of nudity, bestiality, and degrading public figure personalities while these images would have been protected by freedom of expression and categorised as political satire, war photography and visual art.[13] The documentary "suggests there is an urgent need for a parallel investigation into the "international regulatory hand-off," an abdication of responsibility by Big Tech companies, which displace regulatory oversight onto under-paid third-party workforces in developing countries." [14] It has now been reported that while tech companies are eager to grow and expand, they actually fail to address the disastrous effects of their services fomenting hate, discord, and violence.[15]

Lastly, there are **end-users** who, on the one hand, generate content (posts, comments, etc.), and on the other hand, are recipients of content and can therefore flag unwanted content. In addition, platforms such as Reddit, Discord, and Facebook groups rely on volunteer moderators to manage their communities. These volunteers benefit from administrative power over the communities they moderate, such as setting the rules, removing content, and banning people.[16] However, their power only extends to their communities, and they are not competent for other communities or platform-level decisions. They benefit however from a more privileged position to negotiate with the platform on behalf of their communities.[17] This is also called decentralised or self-moderation (see Section 4.1).

[11] PBS, 'The Cleaners', <https://www.pbs.org/independentlens/documentaries/the-cleaners/>, accessed 20 January 2023.

[12] Roberts (n 6); Cambridge Consultants, 'Use of AI in Online Content Moderation' (2019) <https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation> accessed 20 February 2023. For further information content moderator wellness please consult: Steiger and others (n 8).

[13] Lisa Parks, 'Dirty Data: Content Moderation, Regulatory Outsourcing, and The Cleaners' (2019) 73 Film Quarterly 11.

[14] Parks (n 13).

[15] Bryan Bishop, 'The Cleaners Is a Riveting Documentary about How Social Media Might Be Ruining the World' (*The Verge*, 21 January 2018) <https://www.theverge.com/2018/1/21/16916380/sundance-2018-the-cleaners-movie-review-facebook-google-twitter> accessed 9 February 2023.

[16] Joseph Seering and others, 'Moderator Engagement and Community Development in the Age of Algorithms' (2019) 21 New Media & Society 1417.

[17] J Nathan Matias, 'The Civic Labor of Volunteer Moderators Online' (2019) 5 Social Media + Society 2056305119836778.

Opaqueness of the content moderation infrastructure

There are many questions related to the opaqueness of the content moderation infrastructure while it is a complex subject with conflicting rights leading to complex processes design and implementation.[18] As already mentioned, the opaqueness of who exactly, and with what skills, moderates the content is a pressing matter. It has been found that large online platforms often misuse the outsourcing model. To illustrate, many moderators who worked on TikTok content through Majorel, an outsourcing firm in Luxembourg, described experiences of severe psychological distress as a result of their jobs.[19] Another example, according to the press, Facebook has constructed 'a vast infrastructure to keep toxic material off its platform'.[20] Since 2012, the company has hired at least 10 consulting and staffing firms globally to sift through its posts, along with a wider web of subcontractors.[21] The outsourcing took place in regions where the company did not have offices or language expertise. This brought fatal results in the Rohingya crisis.[22] A Reuters investigation showed that the company had dedicated few resources to human content moderators who would understand the local language.[23] Without such capacity, deleting hate speech content is simply impossible. Another controversy has recently aroused because of the lawsuit by a former content moderator employed by Facebook's flagship outsourcing firm in Africa - Sama, alleging severe mental health trauma due to work, as well as other labour violations. Facebook whistle-blower Daniel Motaung has formally launched his case against Facebook and Sama.[24]

In addition, the various layers of the internet are no longer distinguishable, and content moderation is composed of different interfaces and infrastructure layers.[23] The policy

[18] Konstantinos Komaitis, 'Infrastructure And Content Moderation: Challenges And Opportunities' (*Techdirt*, 4 October 2021) <https://www.techdirt.com/2021/10/04/infrastructure-content-moderation-challenges-opportunities/> accessed 5 March 2023.

[19] Rosie Bradbury Al-Waheidi Majd, 'A Factory Line of Terrors: TikTok's African Content Moderators Complain They Were Treated like Robots, Reviewing Videos of Suicide and Animal Cruelty for Less than $3 an Hour.' (*Business Insider*) <https://www.businessinsider.com/tiktoks-african-factory-line-of-terrors-2022-7> accessed 23 February 2023.

[20] Adam Satariano and Mike Isaac, 'The Silent Partner Cleaning Up Facebook for $500 Million a Year' *The New York Times* (31 August 2021) <https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html> accessed 22 February 2023.

[21] Satariano and Isaac (n 20).

[22] Tom Miles, 'U.N. Investigators Cite Facebook Role in Myanmar Crisis' Reuters (12 March 2018) <https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN> accessed 17 March 2023; Amnesty International, 'Myanmar: Facebook's Systems Promoted Violence against Rohingya; Meta Owes Reparations – New Report' (Amnesty International, 29 September 2022) <https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/> accessed 17 March 2023.

[23] 'Why Facebook Is Losing the War on Hate Speech in Myanmar' *Reuters* (15 August 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 23 February 2023.

[24] 'Sama Exploit Facebook Moderators and Call It "Ethical". Help Us Stop Them' (*Foxglove*) <https://www.foxglove.org.uk/campaigns/sama-bcorp/> accessed 23 February 2023.

discussions and research efforts are not so much focused on content moderation infrastructures. Hence, ensuring that the role and responsibilities of infrastructure providers are appropriately scoped is necessary for having a complete picture of the content moderation landscape and challenges. To sum up, a lot of questions, such as the ones below, arise when dealing with AI systems in content moderation: Who gets the data from whom?; Where is the data stored?; Is there an oversight of how the third party moderates the content?; On which dataset is the AI model trained?; Is it the dataset of the company itself or previous datasets owned by the sub-contracted third party?; and, finally, who is liable for workers' mental health issues from reviewing the posts?

The scale of content moderation

To better illustrate the scale of content moderation on major social media platforms, only in the third quarter of 2021, Facebook "took action" on 34.7 million pieces of "adult nudity and sexual activity content", 9.2 million pieces of "bullying and harassment content", 20.9 million pieces of "child sexual exploitation", 22.3 million pieces of hate speech content and 13.6 million pieces of "violence and incitement content". It also took action on 1.8 billion fake accounts.[25] Between April and June 2021, Youtube removed 4.1 million channels and 1 billion comments.[26] Between July and December 2020, Twitter "actioned" on 3.5 million accounts, suspended 1 million accounts, and removed 4.5 million pieces of content.[27] These numbers only represent cases where platforms acted. The overall number of decisions including those where no action was taken is of course much higher. The scale at which these platforms operate means mistakes in enforcing any rule is inevitable: it will always be possible to find examples of both false positives and false negatives.[28] The challenge for platforms is exactly when, how, and why to intervene.[29]

The grounds for content moderation

Importantly, some content moderation decisions - mainly content removals - are required by the EU law, while others are performed voluntarily by platforms.

[25] 'Community Standards Enforcement | Transparency Center' <https://transparency.fb.com/data/community-standards-enforcement/> accessed 1 December 2021.
[26] 'YouTube Community Guidelines Enforcement – Google Transparency Report' <https://transparencyreport.google.com/youtube-policy/removals?hl=en> accessed 1 December 2021.
[27] 'Rules Enforcement - Twitter Transparency Center' <https://transparency.twitter.com/en/reports/rules-enforcement.html> accessed 1 December 2021.
[28] Evelyn Douek, 'Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3679607> accessed 1 December 2021.
[29] Tarleton Gillespie, *Custodians of the internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) <http://www.degruyter.com/document/doi/10.12987/9780300235029/html> accessed 1 December 2021.

Legally required removals are shaped by content moderation legislations detailing what obligations are foreseen for what type of illegal content.[30] Platforms operating under legal frameworks like the US Digital Millennium Copyright Act (DMCA) or the EU's eCommerce Directive and soon the Digital Services Act (DSA) typically mandate a "notice-and-action" system. "Notice and action" is an umbrella term for a range of mechanisms designed to eliminate illegal content from the internet. According to the European Commission, "the notice and action procedures are those followed by the intermediary internet providers for the purpose of combating illegal content upon receipt of notification. The intermediary may, for example, take down illegal content, block it, or request that it be voluntarily taken down by the persons who posted it online."[31]

Then, platforms' voluntary content removals are based on their own set of rules: Community Standards/Guidelines and Terms of Service (ToS), which often include platform operators' own moral beliefs or social norms.[32] Not being mandated by the law, this basis for content moderation decisions has in principle no territorial limitation and no other remedy than the ones offered by the company.[33] These grounds for removal will only apply in cases where the national or European Union laws have not foreseen the illegality of the subject matter. In other words, they complement the removal grounds in the law and are allowed based on the freedom to conduct business of the company. Practically speaking, if the hosting platform is dedicating its space to cat content, it can refuse to have other animal content on its services based on its terms of service.

Moreover, platforms moderate content which belongs to a wide range of categories, including terrorism, graphic violence, toxic speech (hate speech, harassment, and bullying), sexual content, child abuse, and spam/fake account detection. Clearly, these types of content are fundamentally different, not just in terms of their illegality, but also in their characteristics and the gravity of their consequences. It is crucial to recognise that different types of content

[30] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 205395171989794. For the EU, this is typically the TERREG regulation, the CDSM, the CSAM interim regulation,...

[31] European Commission, /* COM/2011/0942 final - 2012/ () */ COMMISSION COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A coherent framework for building trust in the Digital Single Market for e-commerce and online services.

[32] 'Terms of service serves a legal contract between the platform and the users that spells out each party's obligations, liabilities, and other disclaimers, often written in an attempt to avoid future litigation. (...) Community guidelines, on the other hand, often use plain language that explains platforms' expectations of proper user behavior." Jialun 'Aaron' Jiang, 'Toward a Multi-Stakeholder Perspective for Improving Online Content Moderation (Partial PhD in Philosophy)' (Department of Information Science, Faculty of the Graduate School of the University of Colorado 2020).

[33] Rocco Bellanova and Marieke de Goede, 'Co-Producing Security: Platform Content Moderation and European Security Integration' (2022) 60 JCMS: Journal of Common Market Studies 1316.

moderation are fundamentally different, and there is no "one size fits all" solution that may be appropriate in every case. Illegal or potentially problematic content ranges from content that is illegal everywhere (e.g. child sexual abuse material) to content that is legal but potentially harmful (such as disinformation).

Content moderation is a delicate exercise as it is at the crossroad where several fundamental rights meet and where societal concerns arise. Content moderation is a 'powerful mechanism of control'.[34] It evolves through confrontation and cooperation between private companies and public authorities in an international context in light of the internet scale.[35] Therefore it raises the question of who decides, how, and what are the benefits from it. There is a growing body of literature on platform governance analysing the moving power relations between the private actors, including internet and information technology (IT) companies, social media platforms, and public authorities. Through time, the internet has become a prime public space pushing to move from platforms regulating increasingly social and political life towards public authorities aiming to regulate platforms to stay in some sort of control and establish safeguards and boundaries.[36] Some have studied how content moderation can constitute a grip, a policy lever for a public authority to get some control of the increasingly powerful tech actors.[37]

3.1.2 What is algorithmic content moderation?

AI4Media takes a particular interest in the use of automated means for assisting the media sector. This section will therefore devote some focus on AI systems used in content moderation efforts.

Enormous amounts of content are uploaded and circulated on the internet every day, far outpacing any intermediary's ability to have humans analyse content before it is uploaded. Many platforms have therefore turned to automated processes to assist in the detection and analysis of illegal or problematic content, including disinformation, hate speech, and terrorist propaganda.[38] Automated tools bring advantage in terms of scale, cost savings, and speedier decisions.[39] They also promise to relieve workers from the psychological trauma that comes with

[34] Roberts (n 6).
[35] Kyle Langvardt, 'Regulating Online Content Moderation' (2018) 106 The Georgetown Law Journal <https://www.law.georgetown.edu/georgetown-law-journal/in-print/volume-106/volume-106-issue-5-june-2018/regulating-online-content-moderation/> accessed 10 February 2023.
[36] Bellanova and de Goede (n 33).
[37] Robert Gorwa, 'What Is Platform Governance?' (2019) 22 Information, Communication & Society 854.
[38] Emma Llansó and others, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' 30.
[39] Lidia Dutkiewicz and Noémie Krack, 'How to Notice without Looking: The "algorithmization" of Terrorist Content Moderation in the Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online [Part II] - CITIP Blog' <https://www.law.kuleuven.be/citip/blog/how-to-notice-without-looking-the-algorithmization-of-terrorist-content-moderation-in-the-proposal-for-a-regulation-on-preventing-the-dissemination-of-terrorist-content-online-part-ii/> accessed 16 November 2022.

content moderation (see further information in Section 3.1.3). Overall, algorithmic systems have become essential tools for scale content moderation on platforms.[40]

Gorwa et al. define algorithmic (commercial) content moderation as "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g., removal, geoblocking, and account takedown)."[41] AI content moderation systems, therefore, lead to decisions. A distinction must be made between simple filters detecting specific content based on predefined rules and AI systems making a specific decision in relation to such content. Automation can therefore be used in different phases of the content moderation process: i) proactive detection of potentially problematic content, ii) the automated evaluation, or iii) the enforcement of a decision to remove, demonetize, amplify, or prioritise content. AI in content moderation is, therefore, a broad concept and can refer to different technologies at different content moderation stages.[42] Indeed, algorithmic content moderation involves a range of techniques from statistics and computer science. Nevertheless, two main systems are used in algorithmic content moderation: matching and predictive systems.

First, "**matching systems use cryptographic or hashing techniques**. A piece of content is transformed into a 'hash' which is a string of data meant to uniquely identify the underlying content."[43] The system then compares the new piece of content with the hash database containing existing and known content.[44] To illustrate, the Global Internet Forum to Counter Terrorism (GIFCT), a hash-sharing database led by Google, Facebook, Twitter, and Microsoft, plays a significant role in fighting extremism online by removing content it qualifies as "terrorism-related content" under its own terms of service. The technical limitations of hash-sharing technology and the GIFCT database were clearly demonstrated in the Christchurch shooting incident in 2019. On 15 March 2019, a terrorist live-streamed on Facebook his attack on a mosque in Christchurch, New Zealand in which he killed more than 50 people. The live video of the shooting went viral around the world and was able to play for 17 minutes before it was taken down. Including the views during the live broadcast, the video was viewed about 4,000 times in total before being removed from Facebook. Within 24 hours, Facebook had blocked 1.2 million copies at the point of upload and deleted another 300,000. Hundreds of thousands of

[40] Terry Flew, Fiona Martin and Nicolas Suzor, 'Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance' (2019) 10 Journal of Digital Media & Policy 33.
[41] Gorwa, Binns and Katzenbach (n 30).
[42] Noémie Krack and others, 'AI in the Belgian Media Landscape. When Fundamental Risks Meet Regulatory Complexities', *Artificial Intelligence and the Law*, vol 13 (Second Revised Edition, Jan De Bruyne and Cedric Vanleenhove (eds), Intersentia 2023) <https://intersentia.com/en/artificial-intelligence-and-the-law-2nd-edition.html>.
[43] Krack and others (n 42).
[44] Giovanni Sartor and Andrea Loreggia, 'Study for the European Parliament on the Impact of Algorithms for Online Content Filtering or Moderation' (2020) <https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)657101> accessed 26 January 2023.

versions were made and subsequently re-uploaded to Facebook, YouTube, and Twitter.[45] Hash-sharing efforts failed mainly because initial images did not match closely enough to any images already in the database. Even if the technology progresses, it is still easy to go around the hashing detection by editing features of the original image or video (for instance, adding a filter, slowing down, or accelerating the images). In the case at hand, there was not enough similar pre-existing content in the database to allow the machine learning system to match mass shooting-related content[46].

The **second** category includes systems that aim to classify new content into one of a number of categories.[47] They are known as **classification systems** or predictive systems. They consist of the analysis of content with training data in order to identify common features in the content.[48] The training data can contain various types of generalised features related to the content, such as blood, nudity, and keywords blacklisted in vocabulary libraries.[49] This category is even more problematic due to the lack of contextualization, as explained below.

The figure below (Figure 1) illustrates a breakdown of notable algorithmic moderation systems.[50]

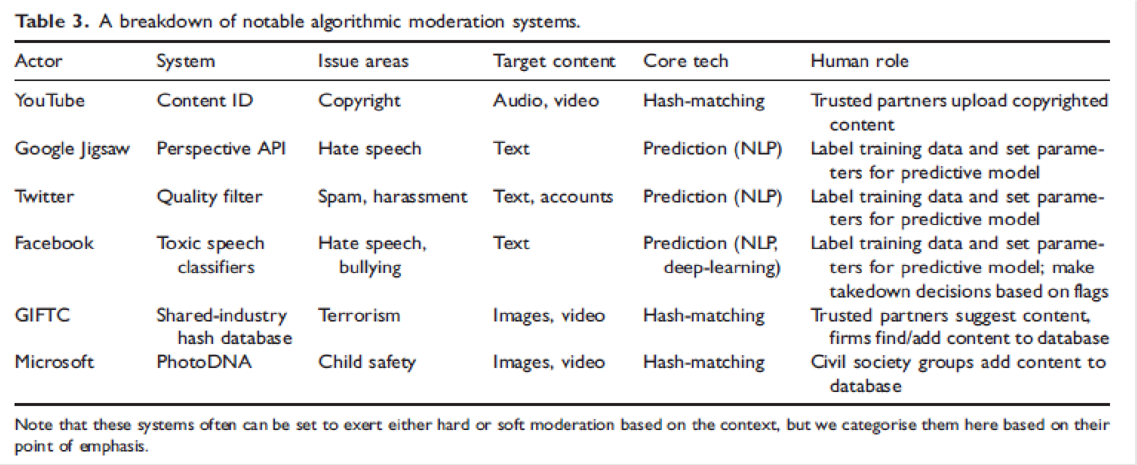


Table 3. A breakdown of notable algorithmic moderation systems.

Actor	System	Issue areas	Target content	Core tech	Human role
YouTube	Content ID	Copyright	Audio, video	Hash-matching	Trusted partners upload copyrighted content
Google Jigsaw	Perspective API	Hate speech	Text	Prediction (NLP)	Label training data and set parameters for predictive model
Twitter	Quality filter	Spam, harassment	Text, accounts	Prediction (NLP)	Label training data and set parameters for predictive model
Facebook	Toxic speech classifiers	Hate speech, bullying	Text	Prediction (NLP, deep-learning)	Label training data and set parameters for predictive model; make takedown decisions based on flags
GIFTC	Shared-industry hash database	Terrorism	Images, video	Hash-matching	Trusted partners suggest content, firms find/add content to database
Microsoft	PhotoDNA	Child safety	Images, video	Hash-matching	Civil society groups add content to database

Note that these systems often can be set to exert either hard or soft moderation based on the context, but we categorise them here based on their point of emphasis.

Figure 1: An overview of algorithmic content moderation practices.
Figure source: Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 205395171989794

[45] Gorwa, Binns and Katzenbach (n 30).
[46] 'Heller - Combating Terrorist-Related Content Through AI and.Pdf' <https://www.ivir.nl/publicaties/download/Hash_sharing_Heller_April_2019.pdf> accessed 2 December 2021.
[47] Gorwa, Binns and Katzenbach (n 30).
[48] Krack and others (n 42).
[49] Heng Sun and Wan Ni, 'Design and Application of an AI-Based Text Content Moderation System' (2022) 2022 Scientific Programming e2576535.
[50] Gorwa, Binns and Katzenbach (n 30).

3.1.3 Challenges and limitations of automation in content moderation

Lack of context

Firstly, it is necessary to emphasise the importance of context**.** Whether a particular post amounts to a violation of law or platforms' Community Standards or Terms of Service often depends on the context that the machine learning system does not recognise.[51] A study on the use of AI tools in hate speech detection points out that these tools are not yet able to understand context, irony, or satire.[52] The best-known example of a lack of contextual differentiation by an online platform's content moderation decision is Facebook's removal of the iconic 'napalm girl' 1972 photo which depicts a young nude girl running from a napalm attack during the Vietnam War.[53] Facebook removed the photo as it breached their Community Standards stating that "while we recognise that this photo is iconic, it's difficult to create a distinction between allowing a photograph of a nude child in one instance and not others."[54] Facebook later reversed its decision and re-instated the photo. Whilst many users would agree that child nudity should be removed from online platforms, this example highlights the importance of context when moderating online content.

Moreover, the lack of contextual interpretation of the terrorist content risks that legal uses of terrorist material (such as for educational, artistic, journalistic, or research purposes, or awareness-raising purposes against terrorist activity) will be deleted. That has happened to the Syrian Archive, a non-profit organisation documenting war crimes committed by terrorist organisations. Its content was repeatedly removed from online platforms, including YouTube, for being "extremist" content and thereby violating platforms' Community Standards and ToS. As a result, the removals can prevent the collection of evidence of war crimes for the International Criminal Court (ICC) or other law enforcement authorities.[55]

In addition, AI systems seem not to be able to react to new contexts, including social, historical and linguistic contexts that they have never encountered in the training or design phase.[56] For instance, an algorithm trained to spot holocaust denial will not be able to detect the Rohingya

[51] Emma Llansó and others (n 38).

[52] Michèle Finck, 'Artificial Intelligence and Online Hate Speech, Centre on Regulation in Europe (CERRE), (2019).

[53] 'Photographer Nick Ut: The Napalm Girl | Buy Photos | AP Images | Collections' <http://www.apimages.com/Collection/Landing/Photographer-Nick-Ut-The-Napalm-Girl-/ebfc0a860aa946ba9e77eb786d46207e> accessed 2 December 2021.

[54] 'Fury over Facebook "Napalm Girl" Censorship' *BBC News* (9 September 2016) <https://www.bbc.com/news/technology-37318031> accessed 2 December 2021.

[55] 'Caught in the Net: The Impact of Extremist Speech Regulations on Human Rights Content' <https://syrianarchive.org/en/lost-found/impact-extremist-human-rights> accessed 1 December 2021.

[56] 'Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis' (*Center for Democracy and Technology*, 20 May 2021) <https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/> accessed 31 January 2023.

genocide. This proves particularly difficult in light of the rising phenomena of an 'algospeak' (see section 4.1.3 on content moderation in metaverse).

Quality, diversity and inclusivity

Second, there is a lack of representative, well-annotated datasets to use for machine learning training. Many tools are trained on labelled datasets that are already publicly available. However, if these datasets do not include examples of speech in different languages and dialects, the resulting tools will not be equipped to analyse these groups' communication. According to the recent Facebook Files, in India, Facebook's single biggest market by audience size, with more than 400 million users, the company's systems were falling short in their effort to crack down on hate speech.[57] The AI models lacked classifiers in the local languages, which need to be trained to detect and remove content such as hate speech. The lack of Hindi and Bengali classifiers means that until 2018 and 2020, respectively, before the company added hate speech to the classifiers, this content was never flagged or actioned. However, these two languages are among India's most popular, spoken collectively by more than 600 million people, according to the country's most recent census in 2011.[58]

The problem with the quality and representation of the training data, especially those in publicly available datasets and databases, is well recognized in the academic literature. As mentioned by Raji and others, "privacy and consent violations in the dataset curation process often disproportionately affect members of marginalised communities. Benchmark dataset curation frequently involves supplementing or highlighting data from a specific population that is underrepresented in the previous dataset".[59] There are a number of studies showing that in the publicly available datasets, certain groups are highly underrepresented. The problem is even more visible when it comes to intersectional identities.[60] To this end, it is likely that using such data could lead to algorithmic results being biased and discriminatory.

Socio-political challenges

Moreover, the process of labelling a dataset for supervised learning typically requires the involvement of multiple human reviewers to evaluate examples and select the appropriate label or to evaluate an automatically applied label. What constitutes "hate speech" or "disinformation" is, however, a socio-political matter and varies across countries and

[57] Rishi Iyengar, 'Facebook Has Language Blind Spots around the World That Allow Hate Speech to Flourish' (*CNN*) <https://www.cnn.com/2021/10/26/tech/facebook-papers-language-hate-speech-international/index.html> accessed 3 December 2021.

[58] Iyengar (n 57)

[59] Inioluwa Deborah Raji and others, 'Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing' [2020] arXiv:2001.00964 [cs] <http://arxiv.org/abs/2001.00964> accessed 27 July 2021.

[60] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' 15.

jurisdictions. The humans applying the label do not agree among themselves on what content merits the label of, for example, “hate speech” or “spam”.

3.1.4 Algorithmic content moderation challenges for freedom of expression and other fundamental rights

Beyond the technical limitations of automated content moderation, the use of automation in content moderation systems raises challenges for freedom of expression and other fundamental rights, which is the main focus of Task 6.1. The use of AI systems may pose a challenge to all fundamental rights, but when it comes to content moderation, the following ones are particularly at stake.

There is always a constant tension between content moderation and freedom of expression. The right to freedom of expression in Europe is enshrined in **Article 10 of the European Convention on Human Rights (ECHR) and Article 11 of the EU Charter on Fundamental Rights (ECFR)**. It includes the right to freely express opinions, views, and ideas and to seek, receive and impart information regardless of frontiers. Freedom of expression is applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference but also to those that offend, shock, or disturb.[61] Users have, moreover, the right to receive and impart information on the internet, in particular, to create, re-use and distribute content using the internet. The right to freedom of expression in Europe has a broad scope of application. It is not limited to citizens or natural persons only. The right protects any expression regardless of its content, its form (any word, picture, image, or action to express an idea, etc.), its speaker, or the type of medium used. There is, however, expression that does not qualify for protection under Article 10 of the ECHR, such as hate speech.[62] It is important to note that the right to freedom of expression is not absolute. In the EU, restrictions can take the form of “formalities, conditions, restrictions or penalties” and are permissible under three conditions. In particular, the restriction must be (1) prescribed by law, (2) introduced for protection of a legitimate aim (e.g., protection of the rights of others), and (3) necessary in a democratic society (proportionate). The rules defining the conditions for lawful interference with expression are addressed to the States and not to private entities.[63]

There is a growing body of literature on the human rights implications of the use of automation by online platforms. The scrutiny over AI tools for content moderation has increased during the pandemic; it showed the overreliance of intermediary services on these tools and the limitations attached to them.[64] There are a number of recognized issues with the application of algorithmic systems in content moderation processes for the end-users.

[61] Handyside v the United Kingdom [1976] ECtHR 5493/72

[62] Erbakan v Turkey [2006] ECtHR 59405/00, Seurot v La France (dec) [2004] ECtHR 57383/00

[63] Kuczerawy (n 3).

[64] ‘Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis’ (n 56).

Freedom of expression

The former UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, has issued a Report on Artificial Intelligence technologies and implications for freedom of expression and the information environment.[65] The Council of Europe has provided several reports, studies, and recommendations that touch on the topic and published in May 2021 its Guidance Note on Content Moderation.[66] The use of automated means by online platforms has also been addressed by the Court of Justice of the European Union (CJEU). In SABAM v. Netlog, the Court held that a filtering system could "undermine freedom of information since that system might not distinguish adequately between lawful and unlawful content, with the result that its introduction could lead to the blocking of lawful communications."[67]

The use of algorithmic systems for detecting particular types of speech and activity will always have so-called false positives (something is wrongly classified as objectionable) and false negatives (the automated tool misses something that should have been classified as objectionable). Both will impact freedom of expression, including the freedom to impart and receive information. Online platforms operate under circumstances in which the cost of over-moderation is low, which makes them set up their content moderation systems to, by default, remove online content or suspend the accounts.[68] These settings would lead to numerous false positives cases. On the other hand, false negatives result in a failure to address hate speech and may create a chilling effect on some individuals' and groups' willingness to participate online.[69] The figure below (Figure 2) illustrates content moderation errors and their consequences.

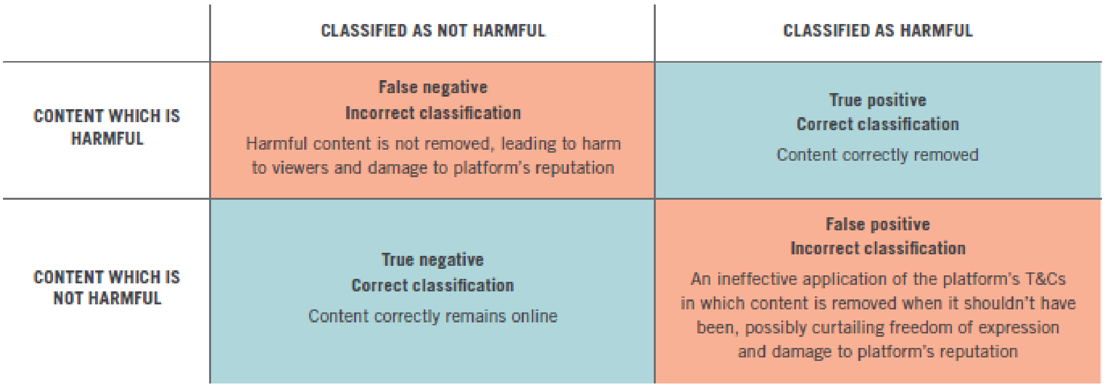
[65] 'OHCHR | Report of the Special Rapporteur to the General Assembly on AI and Its Impact on Freedom of Opinion and Expression' <https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx> accessed 3 December 2021.

[66] 'Guidance Note on Content Moderation' (*Freedom of Expression*) <https://www.coe.int/en/web/freedom-expression/news/-/asset_publisher/thFVuWFiT2Lk/content/guidance-note-on-content-moderation> accessed 3 December 2021.

[67] *Judgment of the Court (Third Chamber), 16 February 2012 Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV, EU:C:2012:85*

[68] 'OHCHR | Report of the Special Rapporteur to the General Assembly on AI and Its Impact on Freedom of Opinion and Expression' (n 65).

[69] Emma Llansó and others (n 38).



	CLASSIFIED AS NOT HARMFUL	CLASSIFIED AS HARMFUL
CONTENT WHICH IS HARMFUL	**False negative** **Incorrect classification** Harmful content is not removed, leading to harm to viewers and damage to platform's reputation	**True positive** **Correct classification** Content correctly removed
CONTENT WHICH IS NOT HARMFUL	**True negative** **Correct classification** Content correctly remains online	**False positive** **Incorrect classification** An ineffective application of the platform's T&Cs in which content is removed when it shouldn't have been, possibly curtailing freedom of expression and damage to platform's reputation

Figure 2: Content moderation errors.
The source of this figure is Ofcom's report on the "Use of AI in online content moderation".

The adverse effects of the use of AI systems for content moderation can directly harm freedom of expression but also indirectly by creating some chilling effect or prior restraints. The AI systems could "filter out parody, irony, content belonging to a legal exemption to intellectual property rights, journalistic or civil society work using illegal content for illustration, negatively impairing legitimate forms of expression and sometimes even privacy"[70].

Right to privacy and protection of personal data (art. 7 ECFR & art. 8 ECHR)

Second, content moderation requires the processing of a range of personal data. Indeed, a range of personal and non-personal data must be stored by the company, such as the username of the individual, the name of the complainant, the justification for the removal of the content, dates and times of uploads and removals, and so on. Furthermore, processing such data may include processing of special categories of data, such as in relation to political opinions, trade union membership, and religious or other beliefs. Such data may only be processed under the General Data Protection Regulation (GDPR) and Convention 108(+) if appropriate safeguards exist in law. Moreover, algorithmic content moderation systems will typically rely on the large-scale processing of user data. This may also involve profiling of users, which is again problematic from a fundamental rights perspective. In this way, the growing reliance on algorithmic systems further encourages the collection and processing of personal data, which poses additional risks to the rights to privacy and freedom of expression.[71] The threats to privacy are a serious concern about the CSAM regulation (see Section 3.2.2.3).

Right to equality and non-discrimination (art. 21 ECFR & Protocol n°12 ECHR)

Third, algorithmic systems have the potential to reproduce and amplify existing discriminations. They can perform badly on data related to underrepresented groups, including racial and ethnic

[70] Krack and others (n 42).
[71] Emma Llansó and others (n 38).

minorities, non-dominant languages, and/or political leanings.[72] This can result in serious risks to freedom of expression for communities and individuals, including illegitimate silencing of their expression and failure to address harms to their communities. As a result, vulnerable groups are the most likely to be disadvantaged by AI content moderation systems.[73] It could result in (deliberate) censorship of a certain group of people. This is either because of the inherent bias of the dataset used to train and test the algorithm, via the platform prioritisation policy which can repress content coming from underrepresented groups or due to lack of proper treatment of individuals' complaints (automatic or not). This all can lead to preventive removal or keyword blacklisting.[74]

Right to a fair trial and effective remedy (art. 47 ECFR & art. 6 ECHR)

Fourth, there is a growing need for redress and accountability for online platforms for making determinations about speech, especially given the enormous scale of speech that is being evaluated. Intermediaries are in a peculiar position as they are better placed to moderate content in light of their technical, financial and human resources. However, "entrusting private stakeholders to take decisions on illegal content puts a great deal of power in their hands without democratic control".[75] As underlined by Aleksandra Kuczerawy, this situation bypasses the protection normally granted by the legal system when the intervention originates from the State and it renders a less visible speech control compared to classic State intervention.[76] When content is removed, it is important that transparency measures make clear the specific reasons why the content was removed. The right to an effective remedy, including complaint, review, and appeal procedures for people whose content has been unjustly removed must be ensured.[77]

Right to property (art. 17 ECFR & art. 1 of Protocol No.1 ECHR)

On the other hand, it is worth mentioning that automated content moderation poses challenges to fundamental rights of third parties, such as the right to property. Owners of intellectual property rights can see their rights heavily infringed online. The AI content moderation systems

[72] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Difering Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 466 (October 2021), 35 pages. https://doi.org/10.1145/3479610

[73] Emma Llansó and others (n 38).

[74] Krack and others (n 42).

[75] Krack and others (n 42); Paul Butcher, 'Disinformation and Democracy: The Home Front in the Information War' (*European Policy Centre*) <https://www.epc.eu/en/publications/Disinformation-and-democracy-The-home-front-in-the-information-war~21c294> accessed 31 January 2023.

[76] Aleksandra Kuczerawy, 'Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?', *Disinformation and digital media as a challenge for democracy*, vol 6 (Cambridge 2020).

[77] Hugo Grotius, *De Jure Belli Ac Pacis. Libri Tres*; Anja Lindroos, 'Addressing Norm Conflicts in a Fragmented Legal System: The Doctrine of Lex Specialis' (2005) 74 Nordic Journal of International Law 27.

of hosting services, if not properly trained, can allow IP infringement material to be hosted on the platform and circulate freely.

3.2 Content moderation landscape

The EU regulatory framework on content moderation is increasingly complex and has been differentiated over the years according to the category of the online platform, the type of content, and the nature of the legal instrument (hard-law, soft-law, or self-regulation). Every legal system must address the hierarchy and relations between norms.[78]

The main elements of the EU regulatory framework include first horizontal rules applicable to all categories of online platforms and all types of content (lex generalis). It includes the e-commerce Directive and the newly adopted Digital Services Act. The AVMSD is a bit peculiar as it is an extra layer of baseline obligations but only for Video-Sharing Platforms. Second, this general framework which can also be called baseline framework is complemented by vertical rules, some lex specialis addressing specific types of content deserving specific attention, rules, and processes. They cover terrorist content, child abuse sexual material, copyright infringing content, racist and xenophobic content, disinformation, and hate speech. Given the various sensitivity or degrees of the illegality of this content, a one size fits all approach would be detrimental to freedom of expression; therefore, specific rules have been adopted. These lex specialis rules are often complemented by self-regulatory initiatives. Lex specialis means that when there is a conflict of laws of equal importance in the hierarchy of norms, the preference/applicability shall be given to the most specific, the one that approaches most nearly to the subject at hand.[79] Therefore lex specialis prevails over lex generalis when there is a conflicting provision or nothing foreseen. These legal concepts have their roots in Roman law. As promptly explained by Hugo Grotius in 1625, weight should be given to that which is regulated more specifically, as indeed, it would seem pointless to apply a more general rule to circumstances already regulated in a more specific manner. This multi-layered content moderation framework will be investigated in the sections below. Figure 3 presents a schematic overview of the lex generalis and lex specialis content moderation frameworks.

[78] Lindroos (n 77).
[79] Lindroos (n 77).

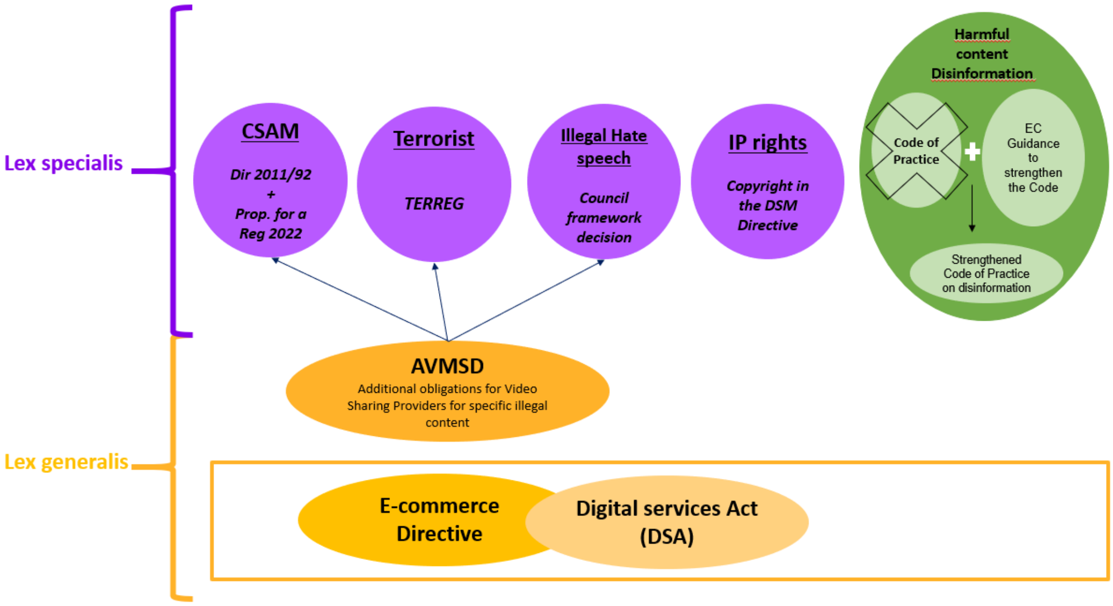

Lex specialis
CSAM
Dir 2011/92
+
Prop. for a
Reg 2022
Terrorist
TERREG
Illegal Hate speech
Council framework decision
IP rights
Copyright in the DSM Directive
Harmful content
Disinformation
Code of Practice
EC Guidance to strengthen the Code
Strengthened Code of Practice on disinformation
AVMSD
Additional obligations for Video Sharing Providers for specific illegal content
Lex generalis
E-commerce Directive
Digital services Act (DSA)

Figure 3: Overview of the EU content moderation landscape[80]

3.2.1 Horizontal rules

3.2.1.1 The e-Commerce Directive

- **Description of the main concepts**

The e-commerce Directive[81], adopted more than 20 years ago, is one of the cornerstones of the Digital Single Market. The goal of this directive was to allow borderless access to digital services across the EU and to harmonise the core aspects of such services, including information requirements and online advertising rules. Until the DSA (see section 3.2.1.3) becomes fully applicable, it remains a 'lex generalis' for the intermediary liability regime. The Directive applies to any kind of illegal or infringing content. It sets out the framework for the liability regime of the so-called intermediary services for third-party (user-generated) content.

The intermediary services are categorised as 'mere conduits', 'caching services', and 'hosting services'. The e-commerce Directive provides for horizontal **liability exemptions**. The idea behind this regime is that imposing liability on platforms for all illegal activity or content related to their services would constitute a considerable burden and prevent e-commerce

[80] Figure 3 is adapting and updating the figure designed in Directorate-General for Internal Policies of the Union (European Parliament) and others, *Online Platforms' Moderation of Illegal Content Online: Laws, Practices and Options for Reform* (Publications Office of the European Union 2020) <https://data.europa.eu/doi/10.2861/831734> accessed 23 January 2023.

[81] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), OJ L 178, 17.7.2000, p. 1–16

development.[82] Each liability exemption is attached to one of the intermediary service categories and is therefore governed by a separate set of conditions enabling the benefits of the exemptions.[83] For instance, under Article 14 of the e-Commerce Directive, **hosting providers** can benefit from a liability exemption provided that: 1) they do not have actual knowledge of illegal activity or information; 2) upon obtaining such knowledge or awareness, they act expeditiously to remove or to disable access to the information.[84] The provider of a hosting service can obtain knowledge about the illicit character of hosted content through his own activities or he could be notified by a private individual to take down the content in question (so-called notice and takedown procedure). As a result, it becomes the provider's task to assess whether the complaint is justified and to make a decision about the illegal or infringing character of the content. The provider can either leave the content on its platform and risk liability for it, or relieve himself of the problem altogether by simply removing the content.[85] The scope of hosting exemptions is quite broad as the case law of the CJEU confirmed its applicability to marketplaces and social media.[86]

Article 15 of the Directive prohibits EU Member States to impose on intermediary service providers a **general obligation to monitor content** that they transmit or store. Member States cannot introduce a general obligation to actively look for facts or circumstances indicating illegal activity. The prohibition of monitoring obligations does not concern monitoring obligations in a specific case.

- **Critical assessment**

Over the span of 20 years of the applicability of the e-commerce Directive, it became clear that the directive presents a series of critical issues. A thorough analysis of the existing problems related to the elimination of illegal content was conducted in the Commission Staff Working Document on Online services, which accompanied the 2012 Communication. Among many of them, the following are worthy to point out.

[82] Christina Angelopoulos and Martin Senftleben, 'An Endless Odyssey? Content Moderation Without General Content Monitoring Obligations' <https://papers.ssrn.com/abstract=3871916> accessed 1 February 2023.
[83] Angelopoulos and Senftleben (n 82).
[84] Aleksandra Kuczerawy, The Power of Positive Thinking: Intermediary Liability and the Effective Enjoyment of the Right to Freedom of Expression, 8 (2017) JIPITEC 226 para 1.
[85] Aleksandra Kuczerawy, 'Safeguards for Freedom of Expression in the Era of Online Gatekeeping' 19.
[86] *Google France SARL and Google Inc v Louis Vuitton Malletier SA (C-236/08), Google France SARL v Viaticum SA and Luteciel SARL (C-237/08) and Google France SARL v Centre national de recherche en relations humaines (CNRRH) SARL and Others (C-238/08)* [2010] ECJ Joined cases C-236/08 to C-238/08; *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECJ Case C-360/10; *L'Oréal SA et autres contre eBay International AG et autres* [2011] Cour de justice Affaire C-324/09.

One fundamental criticism comes from the fact that the e-Commerce Directive grants hosting providers the power to decide which content can remain online and which should be removed. They may be considered private 'gatekeepers', who are able to regulate the behaviour (and speech) of their users. By providing conditional liability exemptions for third parties' illegal content or activities, the States enlist the intermediaries to enforce the public policy objectives (i.e., to remove unlawful content).[87] Tambini[88] calls it 'the first settlement on internet content': huge economic benefits during the internet boom made the governments tackle problems of hate speech, piracy, and harm to children by self-regulation. As the author puts it: "whilst the immense public benefits of free speech over the internet were clear, the framework also permitted a new reality of media freedom to open: net neutrality neutered the ability of networks to control speech, even to protect the public from harm."[89]

Moreover, some criticism refers to the unclear scope of the definitions of online intermediaries, particularly in the case of recent services such as video-sharing sites or social networking sites.

The bulk of the analysis focuses on issues of fragmentation and legal uncertainty. There is a lack of uniform rules for notice and action procedures across the EU. The details of these national obligations vary from member state to member state. This led to a fragmented EU landscape where some member states decided to only obligate hosting service providers to remove content when the notification contains certain information and/or is made by a competent authority.[90] Moreover, Kuczerawy points to the lack of sufficient safeguards to prevent violations of fundamental rights, in particular freedom of expression.[91] The directive does not include provisions, which would provide for effective mechanisms to avoid and/or resolve incorrect removals of content, such as, for example, out-of-court dispute settlement. This lack of safeguards, the authors continue, "leads to over-notification by notifiers, over-removal by providers, and under-assertion of rights by affected users."[92] As noted by Kuczerawy, "the absence of any incentive to conduct a thorough assessment, together with a risk of being held liable, results in a situation where the contested information is often removed or blocked by service providers without giving it a second thought. This leads to situations when legitimate

[87] Aleksandra Kuczerawy (n 84)
[88] Tambini Damian, Media Freedom (Polity 2021)
[89] Tambini (n 89)
[90] Raphaël Gellert and Pieter Wolters, 'The Revision of the European Framework for the Liability and Responsibilities of Hosting Service Providers'.
[91] Aleksandra Kuczerawy, Intermediary Liability and Freedom of Expression in the EU: From Concepts to Safeguards (Intersentia 2018).
[92] Alexandre de Streel and Martin Husovec, 'The E-Commerce Directive as the Cornerstone of the Internal Market' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3637961> accessed 24 January 2023

content, for example, criticism in academic discussion or research, political speech, parody or tribute suffers from such risk-averse behaviour by intermediaries."[93]

Additionally, Article 14 of the e-Commerce does not explicitly impose an obligation on hosting service providers to respond to such notifications (and the subsequent takedowns). The content can therefore be removed before the content providers have a chance to contest the notification, without an opportunity to answer to the allegations of illegality of their content. Several EU countries have introduced 'counter-notification' measures in their national procedures, but it has not become a standard part of the procedure across the EU.[94] Moreover, once a notice has been issued, the hosting provider is expected to react 'expeditiously'. What constitutes an 'expeditious' reaction and what timeframe is foreseen for this action are not specified, and opinions differ as to when this timeframe starts running.[95] Furthermore, the Directive does not envisage that notifications may be sent by bots and fails to incentivise the quality of sent and reviewed notifications.[96]

Some criticism emerged about the prohibition of general monitoring. This concept was not defined in the e-commerce Directive, and the limits of the concept were subject to hot debates throughout the years leading to different interpretations.[97] The question of determining permissible monitoring obligations (specific) and the prohibited ones (general) has been clarified by the CJEU case law. However, the concept seems to vary in the case law depending on the type of content incriminated: copyright infringing content[98] or defamation content.[99] Indeed, the Glawischnig case (defamation) actually has set a turn in the CJEU's constant interpretation of the prohibition of general monitoring obligation on intermediary service providers. It widened the scope of permissible specific monitoring.[100] In this case, the CJEU ruled that a national court can issue an injunction against a hosting provider to detect and remove an illegal message, as well as any equivalent message with an essentially unchanged message without this constituting a general monitoring obligation. This creates uncertainty about the use of AI tools to moderate content as in this case the intermediary had no other option than to monitor all information uploaded by all users, which is contradictory to the previously established CJEU case law on the

[93] Aleksandra Kuczerawy, 'Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative' (2015) 31 Computer Law & Security Review 46.
[94] Kuczerawy, 'Intermediary Liability & Freedom of Expression' (n 93).
[95] Kuczerawy, 'Intermediary Liability & Freedom of Expression' (n 93).
[96] de Streel and Husovec (n 92).
[97] To dive deeper into these different interpretations, we recommend the paper of An Endless Odyssey? Content Moderation Without General Content Monitoring Obligations previously cited in this report.
[98] *L'Oréal SA et autres contre eBay International AG et autres* (n 86); *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* (n 86).
[99] *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECJ Case C-18/18.
[100] Toygar Hasan Oruç, 'The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in Light of Recent Developments: Is It Still Necessary to Maintain It?' (2022) 13 JIPITEC <http://www.jipitec.eu/issues/jipitec-13-3-2022/5555>.

topic.[101] This shows the importance of the case context (defamation) and the evolution of the online environment. Later on, this interpretation was re-iterated but in a copyright case this time and also in an annulment action against the Directive 2019/790/EC on Copyright in the Digital Single Market (CDSM).[102] However, here also it was a case with a notified copyright infringement and where the provider received a sufficiently substantiated notice of the specific infringement or relevant and necessary information regarding the copyright-protected work. These elements must enable the service provider to identify the unlawful content without conducting legal assessment.

The prohibition of general monitoring obligation can lead to paradoxical situations. For instance, a platform carrying out some ex-ante moderation practices to spot illegal content would therefore lose its liability exemption because of this general monitoring prohibition and obligation of conducting a passive role. This passive role would be lost when AI systems are used to proactively search for some content. However, a "refusal or unwillingness to use filters could also be considered as a form of negligence on the part of the intermediary".[103] These contradictory guidelines seem to find a solution in the recent encouragement from the European Commission to intermediary services providers to adopt a more proactive approach to content moderation.[104] Nevertheless, this is not corroborated by the CJEU case law and therefore uncertainty remains despite some clarity brought by the DSA (see section 3.2.1.3). So far, content moderation obligations remain confusing, and have not received a full answer in legal texts or case-law up until now. There is a call to have regulatory explicit clarification stating that "the mere fact that providers use AI-technologies does not automatically preclude the exemption of responsibility".[105] For now, this is very implicitly mentioned in case law, and without a proper framework and clear rules, there are concerns about fundamental rights and respect.

[101] Oruç (n 100).

[102] *Republic of Poland v European Parliament and Council of the European Union* [2022] ECJ Case C-401/19; *Joined Cases C-682/18 and C-683/18: Judgment of the Court (Grand Chamber) of 22 June 2021 (requests for a preliminary ruling from the Bundesgerichtshof — Germany) — Frank Peterson v Google LLC, YouTube LLC, YouTube Inc, Google Germany GmbH (C-682/18) and Elsevier Inc v Cyando AG (C-683/18) (Reference for a preliminary ruling — Intellectual property — Copyright and related rights — Making available and management of a video-sharing platform or a file hosting and -sharing platform — Liability of the operator for infringements of intellectual property rights by users of its platform — Directive 2001/29/EC — Article 3 and Article 8(3) — Concept of 'communication to the public' — Directive 2000/31/EC — Articles 14 and 15 — Conditions for exemption from liability — No knowledge of specific infringements — Notification of such infringements as a condition for obtaining an injunction)* (ECJ).

[103] Emma Llansó and others (n 38).

[104] 'Commission Recommendation (EU) 2018/334 of 1 March 2018 on Measures to Effectively Tackle Illegal Content Online', vol 063 (2018) <http://data.europa.eu/eli/reco/2018/334/oj/eng> accessed 1 February 2023

[105] Alexandre De Streel and others, *Study on Potential Policy Measures to Promote the Uptake and Use of AI in Belgium in Specific Economic Domains* (FPS Economy 2022).

- **Future**

Over time, human capacities of the platforms to prevent and remove illegal content reached their limits. The cost – both financial and emotional – of human content moderators was too high, and more effective automated techniques for identifying and removing illegal content have become available. In light of these developments, the review process of the e-commerce Directive started in 2010, with a public consultation on the future of electronic commerce in the internal market. Later, in May 2015, the Commission announced a plan to assess the role of online platforms in the Communication on a Digital Single Market Strategy for Europe. Finally, in the Communication 'Shaping Europe's Digital Future' in February 2020, the Commission made a commitment to update the horizontal rules that define the responsibilities and obligations of providers of digital services, and online platforms in particular. The Council's Conclusions[106] welcomed the Commission's announcement of a Digital Services Act, emphasised 'the need for clear and harmonised evidence-based rules on responsibilities and accountability for digital services that would guarantee internet intermediaries an appropriate level of legal certainty', and acknowledged 'the need to address the dissemination of hate speech and disinformation online'. It also stressed 'the need for effective and proportionate action against illegal activities and content online (...) whilst ensuring the protection of fundamental rights, in particular the freedom of expression, in an open, free and secure internet.' The work towards re-imagining how digital services work, has started (see section 3.2.1.3 on the DSA).

3.2.1.2 The Audiovisual Media Service Directive (AVMSD)

- **Description of the main concepts**

The Audiovisual Media Service Directive (AVMSD)[107] is the cornerstone of audiovisual media regulation in the EU. The Commission proposed a revision of the old Audiovisual Media Services Directive in May 2016 to include a new approach to online platforms disseminating audiovisual content. The revision of the AVMSD was concluded in November 2018, and Member States had until September 2020 to transpose the AVMSD into their national legislation. The revised AVMSD introduced major changes with regard to the broadening of the scope to include video-sharing platforms (VSPs). Such VSPs – which do not have editorial responsibility over, for example, user-posted videos – have new duties concerning the protection of the general public from:

- content that is illegal under EU law (terrorist content, child sexual abuse material, and racism and xenophobia);

[106] Council conclusions on shaping Europe's digital future Brussels, 9 June 2020 (OR. en) 8711/20.

[107] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities, OJ L 303, 28.11.2018, p. 69–92.

- hate speech based on the illegal grounds mentioned in the EU Charter of Fundamental Rights (sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation).[108]

Additionally, they also have to protect minors from content that may impair their physical, mental or moral development.[109] Audiovisual media services that carry any programmes that might be harmful to minors must provide information such as rating systems or symbols that indicate the presence of violence, nudity, and adult language.

Moreover, the AVMSD created an oversight framework, where national authorities were given the responsibility of verifying that VSPs have adopted "appropriate measures" to deal with different types of content.[110] They include flagging systems, effective complaint systems, age verification, and transparency obligations. It is the obligation of Member States to ensure that all video-sharing platform providers under their jurisdiction apply such measures. Those measures shall be practicable and proportionate, taking into account the size of the video-sharing platform service and the nature of the service that is provided.

It is noteworthy that transparency is given a prominent role in the Directive. First, Art. 28b(3)(d) includes a measure of establishing and operating transparent and user-friendly mechanisms for users of a video-sharing platform to report or flag the content. Second, according to Art. 28b(3)(i), one of the measures may include: "establishing and operating transparent, easy-to-use and effective procedures for the handling and resolution of users' complaints to the video-sharing platform provider in relation to the implementation of the measures referred to in points (d) to (h)".

Kuklis explains that this provision serves both the user who complained and the user against whose content the complaint was directed.[111] This provision is thus potentially a useful tool in protecting the rights of users, especially those who are actively uploading content. The user whose content is taken down by a platform provider usually receives only a generic explanation of the reasons why it happened.[112]

[108] AVMSD, Article 28b (1b) and (1c).
[109] AVMSD, Article 28b (1a).
[110] 'Regulating Content Moderation in Europe beyond the AVMSD' (*Media@LSE*, 25 February 2020) <https://blogs.lse.ac.uk/medialse/2020/02/25/regulating-content-moderation-in-europe-beyond-the-avmsd/> accessed 28 February 2023.
[111] Lubos Kuklis, 'Media Regulation at a Distance: Video-Sharing Platforms in AVMS Directive and the Future of Content Regulation'.
[112] Kuklis (n 111).

- **Critical assessment**

The AVMSD has a very narrow scope of application as it only applies to platforms or to a dissociable section of the platform (service) where the 'principal purpose' or 'essential functionality' is to provide programmes and/or user-generated video content where the service does not have editorial responsibility. An important limiting factor to the AVMSD is that only video content is covered. A service with essentially text or images is excluded.[113] Moreover, the AVMSD only covers services to the extent that they are offered to the general public. Technical internet services, search, online storage, online marketplaces/app stores, and porn publishers are not covered by the Directive.[114]

As provided by Kuklis, from the beginning, one of the most controversial questions was whether VSP regulation in AVMSD would also cover video content on social media.[115] It can be argued that YouTube, TikTok, and all adult VSPs become VSP providers due to the principal purpose of services, while Vimeo or Facebook's Watch Section can be identified as VSPs whose dissociable section of principal service is video sharing.[116] If this assessment cannot be made, then it should be assessed whether the provision of user-generated videos (UGV) or programmes is an "essential functionality" of the service of an online intermediary. To determine this, the EC has identified four main indicators in its Guidelines.[117] They are not legally binding and do not provide uniformity of interpretation.

Another limitation of the AVMSD is that it requires protection from content that is illegal because disseminating it constitutes a crime at the Union level. This means its scope is narrowed to specific very serious crimes (such as terrorism and sexual exploitation).

Importantly, the AVMSD concerns the organisation of the content and not the content itself. The extension of the scope of AVMSD to the regulation of video-sharing platforms is the first time that EU legislation has addressed specific content regulation on any kind of digital platform. For Kuklis, the inclusion of video-sharing platforms in the new version of the Audiovisual Media Service Directive, 'brings a fundamentally new approach to the content regulation as such.'[118] Other instruments predominantly focus on taking down illegal or harmful content and create strong incentives for the platforms for content removal. Thus, the AVMSD is the first legal instrument that provides a catalogue of both procedural (for example, providing for complaint

[113] "Overlaps - Services and Harms in Scope: A Comparison between Recent Initiatives Targeting Digital Services" (*CERRE*February 10, 2023) <https://cerre.eu/publications/overlaps-services-and-harms-in-scope/> accessed March 7, 2023

[114] Overlaps - Services and Harms in Scope (n 113)

[115] Kuklis (n 111).

[116] Oruç (n 100).

[117] Communication from the Commission Guidelines on the practical application of the essential functionality criterion of the definition of a 'video-sharing platform service' under the Audiovisual Media Services Directive 2020/C 223/02, OJ C 223, 7.7.2020, p. 3–9. Not Zotero

[118] Kuklis (n 111).

and redress mechanisms) and technical (for example, age verification and parental control systems) measures to be implemented by the VSPs. Such safeguards ensure a proper oversight of VSP content regulation activities and protect users' right to freedom of expression against potential excess by private actors.

- **Future**

The provisions for VSPs are in a minimum harmonisation regime, which means that member states may choose to impose measures that are more detailed or even stricter than the ones in the Directive. This led to a divergent application of the AVMSD rules. Moreover, the instruments like the AVMSD only cover the dissemination of certain content on certain types of services - only video-sharing platforms and only as regards audiovisual terrorist content or hate speech. As provided by the Impact Assessment to the Digital Services Act, 'while all sector-specific legislative initiatives fulfil their aim to tackle the specific issues, important gaps remain on a horizontal level'.[119] This is why it became necessary to adopt fully-fledged rules applicable to all illegal content in the EU (see Section 3.2.1.3 on the DSA).

3.2.1.3 The Digital Services Act (DSA)

- **Description of the main concepts**

The Digital Services Act (DSA)[120] entered into force on 16 November 2022. The text sets up new due diligence obligations for intermediary services providers and revises/replaces for some part the 20-year-old e-commerce Directive in a "REFIT"[121] exercise by the EC. The scope of the regulation is quite broad and contains a detailed procedural framework. The DSA rules apply to categories of online intermediary services according to their role, size, and impact on the online ecosystem. Online intermediary services such as online marketplaces, app stores, collaborative economy platforms, search engines, and social media platforms will have to comply with a range of obligations to ensure transparency, accountability, and responsibility for their actions.

The DSA maintains the liability rules for providers of intermediary services set out in the e-Commerce Directive – by now established as a foundation of the online sphere. According to Wilman, this approach has been chosen because of the legal certainty that these rules provide as well as their importance for the protection of fundamental rights.[122] By limiting the liability risks that providers face, the rules limit the incentives for providers to remove their users'

[119] Impact assessment of the Digital Services Act, Brussels, 15.12.2020, SWD(2020) 348 final, PART 1/2
[120] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) 2022 (OJ L).
[121] European Commission, 'REFIT – Making EU Law Simpler, Less Costly and Future Proof' <https://commission.europa.eu/law/law-making-process/evaluating-and-improving-existing-laws/refit-making-eu-law-simpler-less-costly-and-future-proof_en> accessed 17 March 2023.
[122] F Wilman, 'THE DIGITAL SERVICES ACT (DSA): AN OVERVIEW'.

information 'just to be sure' in case of doubt about its legality, or to constantly monitor the information transmitted and stored for their users. Such removal and monitoring could have negative consequences for users' fundamental rights (in particular, freedom of expression and information and the rights to privacy and protection of personal data). [123]

Providers of intermediary services, namely mere conduit, caching and hosting services can thus still rely on the relevant liability exemptions, under essentially the same conditions as before. The DSA contains, however, certain clarifications, often based on the case law of the Court of Justice of the EU (CJEU), but their exact scope is beyond the scope of this report.

However, other rules of the e-Commerce Directive were revised. The main aims of the new rules are to:

- establish a horizontal framework for regulatory oversight, accountability and transparency of the online space;
- improve the mechanisms for the removal of illegal content and for the effective protection of users' fundamental rights online, including the freedom of speech;
- propose rules to ensure greater accountability on how platforms moderate content, on advertising and on algorithmic processes;
- provide users with possibilities to challenge the platforms' decisions to remove or label content;
- impose new obligations on very large online platforms (VLOPs) to assess the risks their systems pose and to develop appropriate risk management tools to protect the integrity of their services against the use of manipulative techniques;
- clarify responsibilities and accountability for online platforms and to provide new powers to scrutinize how platforms work, including by facilitating access by researchers to key platform data.

Definition of content moderation

Importantly, the DSA provides the first legal definition of content moderation. Article 3(t) defines 'content moderation' as the activities, whether automated or not, taken by providers of intermediary services, that are aimed, in particular, at detecting, identifying, and addressing illegal content or information incompatible with their terms and conditions. These activities include measures that affect the availability, visibility, and accessibility of that illegal content or information, such as demotion, demonetisation, disabling of access to, or removal, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account. Most importantly, this broad definition of content moderation includes remedies that go beyond content removal. Content moderation is explicitly defined to include not just content removals (takedowns) or account suspension, but also demonetisation and visibility restrictions.

[123] Wilman (n 122).

Content moderation framework

The DSA contains many new provisions aimed at improving content moderation and better tackling illegal content disseminated through intermediary services. In addition, the DSA is providing specific provisions on AI and content moderation. This is a premiere of explicit AI media applications in content moderation.[124]

Clarification of the content and scope of the national orders

The first group consists of rules on orders issued by national judicial or administrative authorities requiring providers of intermediary services to act against certain specific illegal content or to provide information about certain specific users necessary to establish their compliance with the law. Importantly, the DSA does not harmonise what content or behaviour is illegal. According to Art. 3 (h) 'illegal content' means "any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law." This means that what is illegal depends on the national law of Member States. To illustrate, in Europe, laws on abortion vary significantly between countries. For example, Malta and Poland have the strictest abortion laws in Europe, allowing none, or almost none exceptions to the general ban.[125] The DSA lays however certain minimum conditions applicable to orders to act against illegal content. Such orders will have to comply with a number of conditions, including a reference to the legal basis under Union or national law for the order, and a statement of reasons explaining why the information is illegal content, also with reference to one or more specific provisions of Union or national law in compliance with Union law. In the context of access to information about abortion, this means that in a country where access to abortion (but also information about abortion) is restricted, national authority could issue an order to remove such content.[126] The question raises whether an order coming from a country with the strictest abortion law would have to be implemented in all EU countries. The DSA requires for the territorial scope of orders to act against illegal content to be clearly set out on the basis of the applicable rules of Union and national law. Moreover, the territorial scope of an order should be limited to what is strictly necessary to achieve its objective.

Transparency and update of T&C

Article 14 DSA on Terms and Conditions lays down two key principles. First, providers of intermediary services must publish in their terms and conditions, in 'clear and unambiguous language', information on any policies, procedures, measures, and tools used for content

[124] Krack and others (n 42).
[125] Aleksandra Kuczerawy and Lidia Dutkiewicz, 'Accessing Information about Abortion: The Role of Online Platforms Under the EU Digital Services Act' [2022] Verfassungsblog <https://verfassungsblog.de/accessing-information-about-abortion/> accessed 20 March 2023.
[126] Kuczerawy and Dutkiewicz (n 125).

moderation, including "algorithmic decision-making" and human review. In other words, online intermediaries are free to decide what kind of content they do not wish to host, even if this content is not actually illegal. They have to, however, make it clear to their users. They also have to inform them of any significant change to the terms and conditions. Second, these rules must be enforced 'in a diligent, objective and proportionate manner', and with due regard to the interests and fundamental rights involved.

Yearly transparency reports

The DSA now imposes an obligation to draft yearly transparency reports. Initially, art. 23 of the proposal only targeted online platforms but now Article 15 of DSA widens the scope of the obligation by also including intermediary services.[127] These documents will have to, in a clear, easily comprehensible manner, report on any content moderation that they engaged in during the relevant period. This includes the number of orders received from Member States (MS), the number of notices organised by type of illegal content, by trusted flaggers, the number of complaints, and so forth. Additionally, they will also have to report on the use of automated means for content moderation, including a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error, and what safeguards were adopted. These reports will definitely be a mine of information for researchers and civil society, if correctly implemented and drafted. Some big platforms have already started setting up transparency hubs about their services preparing for the implementation of this obligation. This is notably the case of the Meta Transparency Center. However, the accessibility and clarity of these reports will be key to assessing the success of achieving more transparency. The DSA also provides for supplementary elements to report for VLOPs and VLOSEs (Very Large Online Search Engines) such as the human resources dedicated to content moderation, their training, languages expertise, and accuracy indicators per official EU languages (Article 42, para. 2, a).

Notice and action harmonisation

The DSA establishes a notice-and-action framework for content moderation. This mechanism allows users to report the presence of (allegedly) illegal content to the service provider concerned. Article 16 adds additional obligations applicable to providers of hosting services, including online platforms. Providers of hosting services shall put mechanisms in place to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content. The provider is only expected to act if the notice is sufficiently precise and adequately substantiated and the illegality is clear, in that it can be established without a detailed legal examination.[128] The provider shall, without undue delay, notify that individual or entity of its decision in respect of the information to which the notice relates, providing information on the possibilities for redress in respect of

[127] Krack and others (n 42)
[128] Wilman (n 122).

that decision.[129] Moreover, they must take their decisions in a timely, diligent, non-arbitrary, and objective manner. Interestingly, Article 14(6) indicates that providers of hosting services might make use of automated means to make decisions about the notices. When confirming receipt of the notification of a notice they must provide information on such use.[130]

Statement of reason for content moderation decisions

Crucially, Article 17 requires that providers of hosting services provide a clear and specific statement of reasons to any affected recipients of the service on: (a) any restrictions of the visibility of specific items of information provided by the recipient of the service, including removal of content, disabling access to content, or demoting content; (b) suspension, termination or other restriction of monetary payments; (c) suspension or termination of the provision of the service in whole or in part; (d) suspension or termination of the recipient of the service's account, on the ground that the information provided by the recipient of the service is illegal content or incompatible with their terms and conditions. Article 17 lists the information which must be included in such a statement. In particular, where applicable, information on the use of automated means in taking the decision, including information on whether the decision was taken in respect of content detected or identified using automated means.[131]

Three routes for redress about content moderation decisions

The DSA offers three different redress routes that can be used in sequence or separately.[132] Online platforms must put in place an internal complaint-handling system for managing the complaints against a decision taken against information provided/uploaded by a recipient of their services. The decision on the complaint must not be solely taken based on automated means. Article 17 DSA's notification duty does contain two exceptions. First, it does not apply to moderation actions taken in response to removal orders by public authorities, under Article 9 DSA. Second, Article 17(1) DSA exempts content moderation actions affecting 'deceptive high-volume commercial content'.

Other provisions of the DSA require that the content moderation decisions are open to appeals through out-of-court dispute settlement (Article 21). This provides an additional appeal mechanism option against a content moderation decision. The idea behind this, is to fasten the complaint process as for many MS, the judicial journey of a case can take several years, which is out of step compared to the immediacy of online content. "This route of redress can be used as a follow-up, a form of second instance to complaints that have not reached a satisfactory outcome through the internal complaint-handling system. It could also be a self-standing

[129] Art. 16(5) DSA
[130] Art. 16(6) DSA.
[131] Art. 17(3)(c) DSA.
[132] Aleksandra Kuczerawy, 'Remedying Overremoval: The Three-Tiered Approach of the DSA' [2022] Verfassungsblog <https://verfassungsblog.de/remedying-overremoval/> accessed 8 March 2023.

mechanism for complaints that have not been subject to review through the internal system".[133] For Leerseen, this framework reflects the basic principles of due process: every sanction – i.e. any deprivation of lawful interests – must be governed by clear and foreseeable rules; must be notified and explained to the affected users; and must be open to appeals.[134]

The DSA also contains public reporting requirements for content moderation actions (e.g. Articles 15, 23, and 42). The judicial option while not investigated by the DSA remains always available. To dive deeper into this three-tiered approach, we recommend the piece of A. Kuczerawy.[135]

Trusted Flaggers

On top of that, providers of online platforms are required to handle notifications submitted by so-called 'trusted flaggers' with priority.[136] DSA sets in art. 22 the conditions and procedure for becoming Trusted Flaggers, bringing legal certainty to the concept and ensuring harmonisation. To get the status they will need to show particular expertise and competence for the purposes of detecting, identifying and notifying illegal content, be independent of any provider of online platforms, carry out their activities to submit notices diligently, accurately and objectively.

Misuse of rights

Finally, providers of online platforms are obliged to suspend users who misuse their services by frequently providing 'manifestly illegal' content.[137] The user concerned must first have been warned and the suspension must remain limited to a reasonable period. The providers covered by this obligation are also required to assess each case individually, in a timely, diligent and objective manner, as well as to clarify their policies in this regard in advance in the terms and conditions.

Systemic risks assessment and mitigation

Another innovation brought by the DSA is the systemic risks[138] assessment and mitigation. VLOPs and VLOSEs have the obligation to self-assess the systemic risks that their services may cause (art. 34). They must assess how the design of their recommender systems and any other relevant algorithmic system influence these risks, including the dissemination of illegal content. They

[133] Kuczerawy, 'Remedying Overremoval' (n 132).

[134] Paddy Leerssen, 'An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation' (2023) 48 Computer Law & Security Review 105790.

[135] Kuczerawy, 'Remedying Overremoval' (n 132).

[136] Art. 22 DSA.

[137] Art. 23 DSA.

[138] Systemic risks are any negative effects for the exercise of fundamental rights linked to family/ private life, freedom of speech, freedom and pluralism of the media, prohibition of discrimination and children' rights, or the intentional manipulation of their services, including through inauthentic use or automated exploitation means, that has a negative effect on the protection of public health, minors, civic discourse, or on electoral processes and public security.

must also adopt mitigation measures for these identified risks. Mitigation measures can include "content moderation personnel, their training and local expertise and the speed and quality of processing notices related to specific types of illegal content and, where appropriate, expeditious removal of or disabling access to the content notified, in particular for the illegal hate speech or cyber violence; as well as adapting any relevant decision-making processes and dedicated resources for content".[139]

Crisis response mechanism

The DSA also introduces a crisis mechanism granting to the Commission some powers in case a crisis occurs. The crisis will be deemed to have occurred when extraordinary circumstances lead to a serious threat to public security or public health in the Union or significant parts thereof (art. 36). The EC will be, in this case, entitled to request to VLOPs and VLOSEs the urgent adoption of specific measures. These measures can include the adaptation of content moderation processes or relevant algorithms and systems and increasing the resources dedicated to content moderation (rec. 91).

In addition to all these aspects, the DSA develops access and explanations requests about aspects of content moderation which will open the door to more transparency and accountability (art. 37, 40, 69).

- **Critical assessment**

As we can see much of the criticism previously directed towards the e-commerce Directive has been addressed in the DSA. However, there are the following remaining issues.

First, as already mentioned, platforms will have to explain in detail their content moderation policies, i.e. why, when, and how they moderate content. However, as noted by Leersen, most, if not all, major platforms already publish detailed T&C policies.[140] Regardless of legislative requirements, the fundamental problem with these documents is that like all contracts, they would never cover all contingencies and will inevitably leave room for interpretation.[141] Second, there are doubts regarding the actual enforcement of Art. 14. As mentioned, platforms will have to apply their content moderation policies in a diligent, objective, and proportionate manner, and with due regard to the interests and fundamental rights involved. Not only do they have to take 'due regard' to fundamental rights in cases of content removal, but also when restricting the availability, visibility, and accessibility of information. What 'due regard' means in this context will have to be challenged in court. It does not say that certain types of content cannot be removed (or blocked). It emphasises, instead, the importance of proper balancing between different fundamental rights and freedoms. It has been pointed out that it is unclear to which

[139] Krack and others (n 42).
[140] Paddy Leerssen 'An End to Shadow Banning?' (134).
[141] Paddy Leerssen 'An End to Shadow Banning?' (134).

extent users will be able to appeal directly to their fundamental rights, e.g., the freedom to receive and impart information under Article 10 of the ECHR, in a complaint procedure against a platform that restricted content. Who would have such a right? Users whose content was restricted or also third parties whose right to receive information could have been affected as a result of a platform's content moderation decision?[142]

Next, the crisis response mechanism grants the Commission leeway in intervening in the content moderation decision of platforms. This power is limited to 'crisis' situations but the definition is quite broad. This broad scope has been criticised by NGOs[143] as they fear having an "overly broad empowerment of the European Commission to unilaterally declare an EU-wide state of emergency (...) would enable far-reaching restrictions of freedom of expression and of the free access to and dissemination of information in the Union".[144] Article19 and the 22 other signatories of the public statement on crisis mechanism, regret that there is no scrutiny granted to the European Parliament (EP) and that the crisis definition does not fulfil the principles of clarity and specificity. They also question whether the EC is the appropriate body for assessing the occurrence of a crisis unilaterally.[145] The EC has only a report obligation to the EP and the Council without any involvement or say foreseen for them. Some time limits and a proportionality assessment are in the provision (art. 36).

In relation to the intersection of DSA with other content moderation vertical regimes, some pointed out how the dynamic between lex generalis and lex specialis will be more complex than it seems with the DSA implementation.[146] For instance, passing from the E-commerce Directive to a DSA Regulation while some of the lex specialis are still Directives and others are regulations raise concern. What if the implementation of the lex specialis, which is a Directive, greatly differs in Member States? This risks to complicate the interplay of the content moderation landscape, which is already fragmented and complex. However, recently we see in the content moderation landscape a shift from Directives towards Regulations for content moderation lex specialis. This is notably the case with the TERREG regulation and the proposed new CSAM regulation (see Sections 3.2.2.1 and 3.2.2.3.). This approach of moving towards Regulation will indeed simplify the interplay between the lex generalis and lex specialis.

[142] Kuczerawy and Dutkiewicz (n 125).

[143] Article19 and 22 other civil society signatories released a public statement on new crisis response mechanisms and other last-minute additions to the DSA.
Article 19, 'EU: Digital Services Act crisis response mechanism must honour human rights', <https://www.article19.org/resources/eu-digital-services-act-crisis-response-must-respect-human-rights/> accessed 7 March 2023

[144] Article19 (n 143).

[145] Article19 (n 143).

[146] João Pedro Quintais and Sebastian Felix Schwemer, 'The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?' (2022) 13 European Journal of Risk Regulation 191.

Recitals 10 and 11 provide that the DSA "should be without prejudice to other acts of Union law regulating the provision of information society services in general, regulating other aspects of the provision of intermediary services in the internal market or specifying and complementing the harmonised rules set out in this Regulation". While for copyright and related rights, the DSA adds that it is without prejudice to Union law on copyright and related rights, which establish specific rules and procedures that should remain unaffected. Authors argue that this does not mean the horizontal rules would not supplement those in the CDSM Directive, especially as it regards notice-and-action or redress mechanisms.[147]

The DSA explanatory memorandum indicated that the DSA would apply only to the extent that the lex specialis do not contain more specific provisions applicable. These leave several scenarios where the DSA would apply:

- When the DSA rules to regulate matters not covered by lex specialis;
- When the lex specialis leaves some room to MS, but the DSA contains specific obligations on the matter. This latest scenario is less bulletproof, but in terms of democratic legitimacy, it makes sense. [148]

J.P. Quintais and F. Schwemer have provided a blueprint for the DSA liability regime and obligations examination to other sector-specific instruments.[149]

- **Future**

The DSA came into force on 16 November 2022, but it will apply in fifteen months or from 1 January 2024, whichever comes later, after entry into force. Operators designated as very large online platforms (VLOPs) and very large online search engines (VLOSEs) will have to comply with stricter obligations from mid-2023. EU Member States will have to appoint Digital Services Coordinators by 17 February 2024. Given this timeline, it is too early to make recommendations. We first need to see how the DSA will be interpreted, applied and enforced. (See Section 6.1 for a high-level recommendation about the future of content moderation.) Moreover, the EC will adopt implementing and delegating acts framing the application of the DSA.

Nonetheless, Husovec points out a few elements that will make 'the DSA a success story'. [150] First, a community of specialised trusted flaggers would timely and precisely notify problematic content. Second, active individuals who would make use of the DSA tools, consumer associations, dispute resolution bodies, content moderation professionals, and content creators. Third, education and literacy about these new tools, both for ordinary citizens and

[147] Quintais and Schwemer (n 146).
[148] Quintais and Schwemer (n 146).
[149] Quintais and Schwemer (n 146).
[150] Martin Husovec, 'Will the DSA work?: On money and effort' [2022] Verfassungsblog <https://verfassungsblog.de/dsa-money-effort/> accessed 17 February 2023.

researchers on how to make use of the new access to platforms' data mechanism in Article 40. And lastly, strong enforcement.[151]

Moreover, as explained in section 3.2.1.3, the scope of application of the DSA leaves some blank spots. For example, it remains to be seen whether and to what extent the DSA is a future-proof piece of legislation allowing services like metaverse to fall within its scope. The DSA is an important new piece of legislation but it will not achieve its promises if the enforcement part is not ensured. Therefore, strong enforcement will be necessary to materialise all the provisions presented in this section. The DSA has a complex enforcement structure with various actors. Providing actual means and resources for a smooth collaboration will be key to avoid a dead end. The policymakers have learned from the experience with the GDPR where the enforcement mechanism appeared not well designed.[152] The hope is that clear work allocation, collaboration mechanisms, and procedures will solve the problem of the past.[153] In addition, in the future attention will have to be paid to the EC competence centre and the supervisory body as they should be further developed into an independent European authority to prevent political influence.[154]

3.2.2 Vertical rules applicable to illegal content and harmful content

There are vertical rules applicable to the different types of illegal content (i.e. terrorist content, child sexual abuse material, racist and xenophobic hate speech, and content infringing on property rights).

3.2.2.1 Terrorist content

Following a series of terrorist attacks in 2015, including the Charlie Hebdo, Bataclan concert hall, Vienna, and the Brussels airport and metro attacks, the EU decided to adopt measures to stop terrorism. The EU's action is limited as criminal law is a Member States' competence, and the EU can only act in the sphere of security and crime via cooperation and coordination measures (art. 6 TFEU). Among measures such as setting up an EU terrorist list, improving information

[151] Husovec (n 150).

[152] The GDPR enforcement architecture relied heavily on the Data Protection Authority of the country of establishment of the company subject to the GDPR complaint. Many big tech companies have their EU seat established in Ireland, hence most of the cases fell on the Irish DPA's desk. Lack of financial and human resources to assess the cases in depth or political pressure not to scare away the big tech platform? The Irish DPA didn't perform well and created a GDPR enforcement issue. Luca Bertuzzi, 'Ireland's Privacy Watchdog Accused of Paralysing GDPR Enforcement' (www.euractiv.com, 13 September 2021) <https://www.euractiv.com/section/data-protection/news/irelands-privacy-watchdog-accused-of-paralysing-gdpr-enforcement/> accessed 17 March 2023.

[153] Eliska Pirkova, 'The EU Digital Services Act Won't Work without Strong Enforcement' (*Access Now*, 9 December 2021) <https://www.accessnow.org/eu-dsa-enforcement/> accessed 8 March 2023.

[154] Alexandra Geese, 'Why the DSA could save us from the rise of authoritarian regimes' [2022] Verfassungsblog <https://verfassungsblog.de/dsa-authoritarianism/> accessed 8 March 2023.

exchange between law enforcement, judicial and intelligence authorities, setting up an EU Counter-Terrorism Coordinator and financial measures, the EU has also intervened in the content moderation area.[155] Indeed, content moderation became part of the EU's vision for European security.[156]

The same year, the EU Internet Forum to counter terrorism online was created. The Forum provides a collaborative environment for governments in the EU, the internet industry, and other partners to discuss and address the misuse of the internet for terrorist purposes.[157] Thanks to the Forum, in 2016 a shared database of hashes was set up and contains hashes about the terrorist content removed from the online platforms. The EU Internet Referral Unit (IRU) set up in 2015 and now embedded in the Europol European Counter-Terrorism Centre is also an active actor in the content moderation collaboration between public and private entities. Its main objective is to 'refer terrorist and violent extremist content to Online Service Providers (OSPs) and to provide support to member states in the context of internet investigations'.[158] It identifies and refers to terrorist pieces of content to OSP. They operate along with other NGOs as trusted flaggers for terrorist content benefitting from a prioritisation status for the notice submitted.

In 2017, the EU adopted the Counter-Terrorism Directive. The Directive obliges Member States to take the necessary measures to ensure the prompt removal of, or with appropriate safeguards block access to, online content constituting a public provocation to commit a terrorist offence; Member States implemented these obligations via two main types of measures: notice-and-takedown measures and criminal measures.[159] However, as the directive addresses Member States, the measures did not target directly the platforms while they are the ones in a better position to address the topic. The Directive was complemented by a voluntary system for tackling terrorism online based on guidelines and recommendations, but it was deemed insufficient to deal with terrorist content online.[160]

[155] European Council, 'The EU's Response to Terrorism' (15 December 2022) <https://www.consilium.europa.eu/en/policies/fight-against-terrorism/> accessed 3 February 2023.

[156] European Commission, 'Security Union: A Counter-Terrorism Agenda and Stronger Europol to Boost the EU's Resilience' (*European Commission - European Commission*, 9 December 2020) <https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2326> accessed 10 February 2023; European Commission, Communication from the Commission (...)on the EU Security Union Strategy 2020 [COM(2020) 605 final].

[157] European Commission, 'European Union Internet Forum (EUIF)' <https://home-affairs.ec.europa.eu/networks/european-union-internet-forum-euif_en> accessed 9 February 2023.

[158] Europol, 'EU IRU Transparency Report 2019' (*Europol*) <https://www.europol.europa.eu/media-press/newsroom/news/eu-iru-transparency-report-2019> accessed 10 February 2023.

[159]Directorate-General for Internal Policies of the Union (European Parliament) and others (n 80).

[160] Flavia Giglio, 'The New Regulation on Addressing the Dissemination of Terrorist Content Online: A Missed Opportunity to Balance Counter-Terrorism and Fundamental Rights?' (*CITIP blog*, 14 September 2021) <https://www.law.kuleuven.be/citip/blog/the-new-regulation-on-addressing-the-dissemination-of-terrorist-content-online/> accessed 3 February 2023.

This is the reason why, already in September 2018, the European Commission submitted a proposal for a Regulation on preventing the dissemination of terrorist content (TERREG).[161] A regulatory shift is operated by choosing a regulation as an instrument. This instrument imposes directly on Hosting Services Providers duties of care and proactive measures to remove terrorist content including by deploying automated detection tools.

It prescribes a removal of terrorist content within one hour after the order was issued by a national competent authority. It also includes rules concerning complaint mechanisms, transparency obligations, and data retention. This proposal was heavily criticised as it would undermine the prohibition of general monitoring obligations contained in the e-commerce Directive. The shift from the reactive notice and action system towards a more active role from providers seemed to shake the entire ecosystem around the traditional application of liability exemptions and safe harbour. Further criticisms were raised about the broad definition of terrorist content, which could encompass legitimate expression protected under international human rights law, the identity of the competent authority to remove terrorist content, the territorial scope of the removal orders, and the like.[162]

In 2019, following the Christchurch tragedy,[163] the European Union Internet Forum (EUIF) agreed on an EU Crisis Protocol[164] which sets a rapid response to contain the viral spread of terrorist and violent extremist content online. This system is only reserved for extraordinary situations. The Crisis Protocol provides a coordinated and rapid reaction for Member States' authorities, Europol, the Global Internet Forum to Counter Terrorism (GIFCT), and online service providers. Cooperation is reinforced in the event of a crisis. For instance, similar to Christchurch, URLs, content, and metadata can be more easily shared in real-time.

[161] Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online A contribution from the European Commission to the Leaders' meeting in Salzburg on 19-20 September 2018 2018 [COM/2018/640 final].

[162] Lidia Dutkiewicz and Noémie Krack, 'All Eyes Riveted on the Trilogue Closed Doors of the Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online [Part I]' (*CITIP blog*, 24 November 2020) <https://www.law.kuleuven.be/citip/blog/all-eyes-riveted-on-the-trilogue-closed-doors-of-the-proposal-for-a-regulation-on-preventing-the-dissemination-of-terrorist-content-online-part-i/> accessed 16 November 2022.

[163] In March 2019, two terrorist attacks against Mosques took place in Christchurch in New Zealand. The attacks were live streamed on Facebook. The videos didn't get immediately reported and taken down, which enabled capture of the images and videos. The footage was reuploaded millions times in various platforms. Users by editing the video outsmarted the content moderation hash database in place. For more information on the event see Kristina Hummel, 'The Christchurch Attacks: Livestream Terror in the Viral Video Age' (*Combating Terrorism Center at West Point*, 18 July 2019) <https://ctc.westpoint.edu/christchurch-attacks-livestream-terror-viral-video-age/> accessed 9 February 2023.

[164] European Commission, 'EU Internet Forum Committed to an EU-Wide Crisis Protocol' (European Commission) <https://ec.europa.eu/commission/presscorner/detail/en/IP_19_6009> accessed 17 March 2023.

The proposal now finished its journey in the EU legislative-making process and was adopted. The Regulation was published at the Official Journal (OJ) in May 2021, entered into force on 6 June 2021, and applies as of 7 June 2022.[165] The next paragraphs will analyse the main concepts in the final text and provide a critical analysis.

- **Description of the main concepts**

Now, a competent authority of a Member State can issue a removal order requiring hosting service providers to remove terrorist content or to disable access to such content in the whole European Union. The competent authority is not necessarily a judicial body. Content referrals sent from either a national competent authority or an EU body such as Europol that the Hosting Service Providers (HSPs) must expeditiously assess. The window time for action upon receipt of an order requires terrorist content to be removed within one hour from the receipt of the removal order and imposes financial penalties for non-compliance. The Regulation grants freedom to hosting service providers on their choice of specific measures to comply with the Regulation. On the condition, however, that these measures are effective in mitigating the risk, proportionate with the technical, financial, and operational capabilities, the number of users of the hosting service provider and the amount of content they provide. The imposition of any requirement leading to a general obligation to monitor or actively seek facts or circumstances indicating illegal activity under Article 15(1) ECD or to use of automated tools by hosting providers is prohibited. In addition, competent authority can also decide that a certain HSP is particularly exposed to terrorist content. They can oblige the hosting service to adopt measures to prevent the dissemination of terrorist content on its services.

- **Critical assessment**

There is a normative tension between the EU security-policy making and the EU's stance as a protector of freedom of expression and free press.[166] Through the TERREG regulation, we can observe how the EU is seeking to have a more active role in "steering and influencing private practices and decisions on content removal".[167] Regulation on preventing the dissemination of terrorist content online has been subject to many controversies. In March 2021, 61 human rights organisations signed a joint letter to the European Parliament calling to vote against the text.[168] The letter points out how the text poses serious threats to freedom of expression and opinion, freedom to access information, the right to privacy, and the rule of law. Indeed, the TERREG text enables MS restrictions on online speech after only a minimal review.

[165] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance) 2021 (OJ L).

[166] Bellanova and de Goede (n 33).

[167] ibid.

[168] European Digital Rights (EDRi), 'Coalition of Human Rights and Journalist Organisations Express Concerns for Free Speech' (*European Digital Rights (EDRi)*) <https://edri.org/our-work/coalition-humn-rights-media-organisations-express-gave-concerns-free-speech/> accessed 9 February 2023.

The 1-hour window for action upon order receipt

The one-hour response deadline has been criticised for being very difficult to meet in practice and for creating incentives for the over-removal of content. Multiple NGOs have also underlined that such removal orders "must be met within this short time period regardless of any legitimate objections platforms or their users may have to the removal of the content specified, and the damage to freedom of expression and access to information may already be irreversible by the time any future appeal process is complete”.[169] In practice, hosting service providers would rather prefer to delete more content, faster, e.g. by installing upload filters to systematically monitor the entirety of the users’ content, to avoid facing financial penalties. This risks negatively affecting fundamental rights, in particular the right to freedom of expression.

Push towards a more proactive role from hosting providers and hence towards algorithmic moderation

The new duties of care contained in the regulation actually push hosting service providers to take proactive measures to ensure compliance with the obligations laid down. Despite the prohibition on general monitoring, “the whole system established under the Regulation including the obligation to remove notified terrorist content and to take specific measures for the protection of the service, seems to give no other option to hosting providers but to take certain proactive measures in practice.”[170] Article 5(2)(a) classifies “appropriate technical and operational measures or capacities, such as appropriate staffing or technical means to identify and expeditiously remove or disable access to terrorist content” as a permissible specific measure which clearly requires *de facto* monitoring of uploaded content in order to identify terrorist content.[171] Proactive measures can only be fulfilled thanks to the help of AI systems and upload filters.[172] The limited time window leaves basically no choice in the means used to comply with the law. The text also provides explicitly the possibility for hosting providers to use automated tools if they consider it to be appropriate and necessary to effectively address the misuse of their service. The choice of measures leaves a considerable margin of appreciation to the private actors.[173]

[169]Article 19, ‘Joint Letter on European Commission Regulation on Online Terrorist Content’ (6 December 2018) 19 <https://www.article19.org/resources/joint-letter-on-european-commission-regulation-on-online-terrorist-content/> accessed 9 February 2023.

[170] Oruç (n 100).

[171] ibid.

[172] Clara Rauchegger and Aleksandra Kuczerawy, ‘Injunctions to Remove Illegal Online Content under the Ecommerce Directive: Glawischnig-Piesczek’ <https://papers.ssrn.com/abstract=3728597> accessed 25 November 2022.

[173] Flavia Giglio, ‘The New Regulation on Addressing the Dissemination of Terrorist Content Online: A Missed Opportunity to Balance Counter-Terrorism and Fundamental Rights?’ (CITIP blog, 14 September 2021) <https://www.law.kuleuven.be/citip/blog/the-new-regulation-on-addressing-the-dissemination-of-terrorist-content-online/> accessed 3 February 2023

Lack of differentiation between the size of the company

The TERREG covers hosting service providers when they disseminate information to the public. It does provide the exception that the regulation does not apply to "material disseminated to the public for educational, journalistic, artistic or research purposes or for the purposes of preventing or countering terrorism, including material which represents an expression of polemic or controversial views in the course of public debate". It further foresees that an "assessment shall determine the true purpose of that dissemination and whether material is disseminated to the public for those purposes". However, no further guidance is given, and it is uncertain how such an assessment should be done.[174] Moreover, all platforms in scope are subject to the same obligations in relation to terrorist content. They all have to adopt measures, no matter the size or reach of the service.

Lack of independent judicial review for takedown orders

TERREG does not impose an independent judicial review for takedown orders. Therefore, civil society fears that the instrument could be abused for politically motivated censorship.[175]

Insufficient transparency

According to some scholars, the TERREG regulation is enabling an "opaque configuration of public-private security collaboration within security".[176] It is a standard practice that online content, such as a tweet, a picture, or a video is referred by EU Internet Referral Units (IRUs) to the relevant platform. Criticism has been leveraged against the IRUs as this cooperation risks undermining the rule of law. This is because referrals of terrorist content can promote content removal via extra-legal channels based on company terms of service. Whilst the Unit does release annual transparency reports, there is no formal oversight of judicial review of the EU IRU's activities.[177] Moreover, datasets extracted from social media platforms can become part of databases such as the one of Europol, and can, under specific conditions, be processed for other purposes. More transparency on the other purposes should be provided especially in light of the growing role of Europol.[178]

- **Future**

Addressing terrorist content is a challenge as it is a moving target. While machine learning models train on historical data, terrorist propaganda changes over time. As shown in an analysis above, TERREG contains a lot of controversial elements. It should be noted that the French

[174] Overlaps - Services and Harms in Scope (n 113)
[175] European Digital Rights (EDRi) (n 168).
[176] Bellanova and de Goede (n 33).
[177] 'THE ONLINE REGULATION SERIES | EUROPEAN UNION (Update) - Tech Against Terrorism' (10 December 2021) <https://www.techagainstterrorism.org/2021/12/10/the-online-regulation-series-european-union-update/, https://www.techagainstterrorism.org/2021/12/10/the-online-regulation-series-european-union-update/> accessed 7 March 2023.
[178] Bellanova and de Goede (n 33).

Constitutional Court struck down the so-called Avia Law[179], questioning the provisions that are also present in TERREG: 1-hour content takedown deadline for removal orders. It, therefore, remains to be seen if TERREG will be challenged by some Member States. If so, we might see the disputes over TERREG's controversial provisions being tackled by the Court of Justice of the EU.

As with other instruments, much will also depend on how Member States interpret and enforce the Regulation. For now, on 27 January 2023, the Commission has decided to send letters of formal notice to 22 Member States[180] for failing to comply with certain obligations from the Regulation on the dissemination of terrorist content online, such as: the requirement to designate the authority or authorities responsible for issuing removal orders and notify the Commission of those authorities; to name a public contact point and to lay down the rules and measures on penalties in case of non-compliance with legal obligations.

3.2.2.2 Copyright-protected content

While online platforms, in other words content-sharing service providers, brought many affirmative opportunities for sharing and creating content without barriers, a large number of daily uploads to these platforms made it harder to assess their lawfulness when it comes to copyright-protected works. Thus, this created tension between rightsholders and online platforms regarding the conditions their works and other subject matter are used and whether they would be able to obtain appropriate remuneration for such use.[181] With the aim of clarifying this uncertainty, endorsing the licensing agreements between platforms and rightsholders, and harmonising certain aspects of EU Member States' copyright legislation, the Directive 2019/790/EC on Copyright in the Digital Single Market (CDSM) came into force in 2019, with the member states transposition deadline of June 7th, 2021. The CDSM is considered lex specialis compared to lex generalis instruments on content moderation and intermediary liability such as the E-Commerce Directive and the newly adopted DSA. Therefore, while the E-Commerce Directive and the DSA are still applicable to issues related to content moderation, the CDSM's specific provisions will be given priority on issues concerning copyright-protected content.

- **Description of the main concepts**

Art. 17 of the CDSM

One of the most debate-provoking provisions of the CDSM has been Art. 17 (ex-Art. 13), which imposes direct liability on online content-sharing service providers (OCSSPs)[182] for copyright-

[179] EDRi, 'French Avia Law Declared Unconstitutional: What Does This Teach Us at EU Level?' (European Digital Rights (EDRi), 24 June 2020) <https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/> accessed 17 March 2023

[180] Belgium, Bulgaria, Czechia, Denmark, Estonia, Ireland, Greece, Spain, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Finland and Sweden.

[181] Recital 61, the Copyright in the Digital Single Market Directive.

[182] Recital 62, CDSM defines the OCSSPs as "services, the main or one of the main purposes of which is to store and enable users to upload and share a large amount of copyright-protected content with the

protected works or other protected subject-matter uploaded by users.[183] Art. 17(1) clarifies that the reasoning behind such direct liability is because under Art. 3 of Directive 2001/29/EC (InfoSoc Directive)[184] OCSSPs perform an act of communication to the public when they give the public access to copyright-protected content. Thus, they are obliged to obtain authorisation from the rightsholders, by concluding licensing agreement or in the form of other methods that would qualify as authorisation. Additionally, under Art. 17(3), it is stated that the OCCSPs performing acts falling under the CDSM would not be able to benefit from the limitation of liability established in Art. 14(1) of the E-Commerce Directive. As Art. 17 is a huge provision with 10 long sub-paragraphs, for the scope of this deliverable, only relevant sub-paragraphs of Art. 17 will be explained and analysed below.

According to Art. 17(4), the platforms could avoid this liability for user-generated content (UGC) if they have: "(a) made best efforts to obtain an authorisation; (b) made, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the rightsholders have provided the service providers with the relevant and necessary information; and in any event; (c) acted expeditiously, upon receiving a sufficiently substantiated notice from the rightsholders to disable access to, or to remove from their websites, the notified works or other subject matter, and made best efforts to prevent their future uploads in accordance with point (b)."

Although this three-tiered standard appears like an ordinary liability provision, apart from imposing the obligation to licence, there are several regimes introduced concerning copyright-protected UGC including so-called notice-and-takedown and notice-and-stay-down. Under the notice-and-takedown process, platforms are required to make 'best efforts' to takedown copyright infringing UGC upon receiving notice from rightsholders. Similar regimes exist in other jurisdictions such as Mexico, New Zealand, Canada, and alike, mostly with the influence of the US Digital Millenium Copyright Act of 1998 (DMCA).[185] DMCA §512[186] provides a safe harbour to intermediaries that comply with the conditions of the notice-and-takedown system, which

purpose of obtaining profit therefrom, either directly or indirectly, by organising it and promoting it in order to attract a larger audience, including by categorising it and using targeted promotion within it." Furthermore, the following providers of services excluded from the definition of OCSSPs: open source software development and sharing platforms, not-for-profit scientific or educational repositories as well as not-for-profit online encyclopedias, as well as electronic communication services, business-to-business cloud services and cloud services, which allow users to upload content for their own use, such as cyberlockers, or online marketplaces the main activity of which is online retail, and not giving access to copyright-protected content.

[183] Art. 17(1), the Copyright in the Digital Single Market Directive.

[184] Art. 3, the Information Society Directive.

[185] Emine Ozge Yildirim and others, 'Freedom to Share: How the Law of Platform Liability Impacts Licensors and Users,' (Creative Commons Medium, 2021), <https://medium.com/creative-commons-we-like-to-share/freedom-to-share-how-the-law-of-platform-liability-impacts-licensors-and-users-84d86adade4e>

[186] 17 U.S.C §512.

includes notice and counter-notice mechanisms. In theory, both DMCA §512 and Art. 17(9) respect the possibility of reinstatement of erroneously removed or blocked materials. However, in practice, the situation differs drastically as will be explained below. When it comes to the notice-and-stay-down process, platforms are obliged to make the 'best efforts' to prevent future uploads of works that have been taken down after notice from rightsholders or had previously been flagged as infringing. Art. 17 differs from the DMCA here, as the current US legal framework does not include such a requirement for platforms to comply with in order to shield their safe harbour status.

Furthermore, Art. 17(7) provides that the above-mentioned requirements should not affect the availability of works that do not infringe copyright and related rights, including works or other subject matter that are covered by an exception or limitation.[187] The legislation also reaffirms, in line with the other instruments on content moderation and the CJEU precedent, that application of this provision must not lead to any general monitoring obligation.[188]

- **Critical assessment**

Upload Filters

Art. 17 does not explicitly mandate or set forth a legal obligation for platforms to utilise automated content recognition technologies or upload filters to detect and takedown infringing content. Nonetheless, Art. 17(1) read together with Art. 17(4) requires acquiring a license or ensuring the unavailability of content for which it could not obtain a license. Additionally, if the platform cannot satisfy the very vague 'best efforts' standard of taking down infringing content and ensuring keeping it down, it can possibly be held liable for infringement.[189] Such a fear of liability has the potential of inducing platforms to use ex-ante upload filters to remove or block content before it even has a chance to be made available to the public. Unfortunately, this is not a futuristic scenario currently, as platforms have started to adopt filtering technologies without being legally mandated to do so. Youtube's Content ID and Facebook's Rights Manager Tool are among the examples of such tools.

One may question why using upload filters or other similar technologies would be such a big issue if it allows platforms to avoid liability and keep providing their services. The short answer would be that automated filtering technologies come with their limitations. Currently, no filtering technology is capable of understanding the context, purpose, or nuance of the use of such work. Thus, they are not able to avoid false positives,[190] and they can potentially remove or block lawful content. This includes not being able to identify exceptions or limitations the

[187] Art. 17(7), the Copyright in the Digital Single Market Directive.
[188] Art. 17(8), the Copyright in the Digital Single Market Directive.
[189] Christophe Geiger and Bernd Justin Jütte, 'Platform Liability under Art. 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match' (2021) 70 GRUR International 517.
[190] ibid.

work might benefit from. Despite Art. 17(7) sets forth that the cooperation between the OCSSPs and the rightsholders should not affect the non-infringing content, there are several examples of lawful content being taken down by upload filters. For instance, in 2012, Youtube's Content ID took down a NASA-uploaded public domain video briefly.[191] The takedown was not mandated by a notice under the DMCA or any other legal instrument; it was simply filtered by Content ID, as it was not able to understand that the content was in the public domain. As a result, public domain or openly licensed content, as well as content shared under parody, quotation, and other exceptions and limitations are all in danger of erroneously being removed ex-ante. Thus, such filtering also has the potential of triggering over-removal and over-blocking of lawful content, which implies freedom of expression concerns that will be expanded upon below.[192]

Reinstatement of Removed Content and Bad Faith Notices

The notice-and-takedown and notice-and-stay-down regimes could be abused by malicious actors with bad faith notices that could result in erroneous removals.[193] Such notices could especially interfere with the enjoyment of benefitting from exceptions and limitations, allowing some actors to profit from the removal. Additionally, according to Bridy and Keller, successful counter-notices to removals or blockings are rare, possibly due to unclarity on whether users actually receive notice of removal or the intimidating nature of the counter-notice process.[194] Despite Art. 17(9) providing that the OCSSPs should put in place effective and expeditious complaint and redress mechanisms for reinstatement of removed or blocked content, the platforms have no sufficient incentive or perhaps clear guidance to do so. Therefore, users do not have adequate and effective ex-ante and ex-post redress mechanisms to safeguard their legitimate uses of works, again leading to the removal or blocking of lawful content.

Freedom of Expression Concerns

As mentioned, the use of upload filters, bad faith notices, and the lack of adequate redress mechanisms could result in erroneous removing, over-removing, or over-blocking, therefore, censoring lawful content. Censorship of this nature significantly hinders the enjoyment of exercising certain fundamental rights, particularly by chilling freedom of expression and information.[195] As the availability of content shrinks and lawful uses are disrupted, users are at the risk of being denied the opportunity to access knowledge, join the public debate, seek creative pursuits dependent on the availability of information, freely share, and let their voices

[191] Mike Masnick, 'How Google's ContentID System Fails at Fair Use & the Public Domain' (*Techdirt,* January 1, 2021) <https://www.techdirt.com/2012/08/08/how-googles-contentid-system-fails-fair-use-public-domain/> accessed February 20, 2023.

[192] Sevra Guzel, 'Article 17 of the CDSM Directive and the Fundamental Rights: Shaping the Future of the Internet,' European Journal of Law Technology Vol.12 No: 1 [2021].

[193] *See* Lenz v. Universal Music Corp., 801 F.3d 1126 (US Court of Appeals, 9th Cir. 2015)

[194] Annemarie Bridy and Daphne Keller, 'U.S. Copyright Office Section 512 Study: Comments in Response to Notice of Inquiry' [2016] SSRN Electronic Journal.

[195] Guzel (n 192)

be heard. Additionally, even though Art. 17(8) reaffirms the ban on general monitoring obligations, filtering a massive amount of content with the threat of liability pushes platforms to conduct quasi-general monitoring. Due to similar fundamental rights concerns, Poland filed an action for annulment of Art. 17 with the CJEU claiming that the Article violates fundamental rights enshrined in the EU Charter of Fundamental Rights, especially Art. 11 on Freedom of expression and information.[196] The action by Poland was dismissed by the CJEU in 2022. In the judgement, while the CJEU recognized that upload filters could result in over-blocking and interference with users' rights, it was stated that Art. 17 provides adequate procedural safeguards to protect the right to freedom of expression and strikes a fair balance between competing interests and rights.[197] The Court also concluded that Art. 17's application must not lead to a general monitoring obligation, as the OCSSPs cannot be required to monitor content to determine its lawfulness.

- **Future**

The future of our communities depends on the availability of knowledge and creativity derived from such knowledge. This requires policymakers and platforms to take steps to preserve content that is made available to the public lawfully, without infringing rightsholders' rights or leaving them in a disadvantageous position. Therefore, several means would aid in achieving this end.

Policymakers should discourage platforms from using preventive upload filter technologies by making clear their responsibilities concerning potentially infringing content. The use of such filters could only be justified in the case of removing or blocking manifestly infringing content. Thus, for allegedly infringing content, the use of upload filters should be prevented as much as possible. Instead, ex-ante human review for allegedly infringing content should be guaranteed.

Regarding the redress process, after content is removed or blocked, a strengthened version of ex-post human review should be ensured to make sure that lawful use of content is not being taken down erroneously or with bad faith notices. Additionally, quick and effective reinstatement mechanisms should be put in place, so that the removed content could be put back online without delayed processes and possibly staying down. This also means that the platforms should be encouraged to put in place mechanisms that would notify the persons whose content was removed with sufficient and simple information on how to contest if the removal was erroneous. So that counter-notice processes could be used more efficiently. For instance, the DSA provides more procedural safeguards concerning redress mechanisms. In theory, OCSSPs are exempted from the general liability of the E-Commerce Directive and the

[196] CJEU, Case C-401/19 Republic of *Poland v. European Parliament and Council of the European Union* [2022].

[197] João Pedro Quintais, 'Between Filters and Fundamental Rights: How the Court of Justice saved Article 17 in C-401/19 - Poland v. Parliament and Council,' (VerfBlog, 2022/5/16), <https://verfassungsblog.de/filters-poland/>

DSA, due to their more disruptive nature. However, according to Quintais and Schwemer, some provisions introduced by the DSA should be horizontally applicable to OCSSPs as well,[198] which are also hosting providers and online platforms. Therefore, further clarification by the legislator should also be provided in this respect.

Lastly, the platforms should be given enough incentives and guidance to adopt preventive measures and policies to safeguard works shared under an exception or limitation, as well as works that are in the public domain or shared under a non-exclusive license. Apart from adopting such policies, policymakers could assist platforms for ex-ante reviews, especially for distinguishing manifestly infringing content from allegedly infringing content. For instance, this could mean that policymakers could create a database consisting of a centralised repository or European Commons of public domain and non-exclusive licensed works, where it is easy for upload filters to recognise whether the content is an infringing or a lawful use.

3.2.2.3 Child sexual abuse material

Child sexual abuse and violence are touching one in five children according to the Council of Europe.[199] According to the EC, 85 million pictures and videos depicting child sexual abuse were reported worldwide in 2021 alone.[200] This frightening number only demonstrated the reported cases of CSAM, and it is the tip of the iceberg. The ever-growing use of social media and the internet by children and teenagers exposes them to greater threats, especially with grooming practices or when their sexual abuse is being recorded, uploaded, or streamed online making it extremely hard to remove completely heal or reconstruct.[201] This content cannot freely circulate on the internet and EU legislation was adopted to impose specific moderation obligations and responsibilities to platforms to deal with it to prevent their spread and harmful impacts. Specific rules at the EU level are needed to ensure a consistent approach to address the matter. Especially in light of the cross-border aspect of it, the long-term physical, psychological, and social harm to victims, and the negative impact on the core values of a modern society relating to the special protection of children and trust in relevant State institutions.[202] Article 24 of the Charter of Fundamental Rights of the European Union, lays down a positive obligation to act with the aim of ensuring the necessary protection of children in line with the UN Convention on the Rights of the Child. Despite the particular seriousness of the crime, any limitation to

[198] Quintais and Schwemer (n 146).

[199] 'The Underwear Rule - Children's Rights - Publi.Coe.Int' (*Children's Rights*) <https://www.coe.int/en/web/children/underwear-rule> accessed 20 January 2023.

[200] 'Fighting Child Sexual Abuse' (*European Commission - European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/IP_22_2976> accessed 20 January 2023.

[201] Hee-Eun Lee and others, 'Detecting Child Sexual Abuse Material: A Comprehensive Survey' (2020) 34 Forensic Science International: Digital Investigation 301022.

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

fundamental rights impacted by the fight against CSAM must be done in accordance with the principle of proportionality. This appears to be a delicate balance difficult to reach when it comes to CSAM regulation.

- **Description of the main concepts**

In 2004, a Council Framework Decision 2004/68/JHA[203] introduced a minimum approximation of Member States' legislation to criminalise the most serious forms of child sexual abuse and exploitation, to extend domestic jurisdiction, and to provide for a minimum of assistance to victims. The decision was completed by several other decisions such as the Council Decision of 29 May 2000 to combat child pornography on the internet[204] and the Decision No 854/2005/EC of the European Parliament and of the Council establishing a multiannual Community Programme on promoting safer use of the internet and new online technologies[205].

Child Sexual Abuse and Exploitation Directive (CSAED)

This 2004/68/JHA decision was repealed and replaced in 2011 when CSAM started to be regulated through EU legislation with the Child Sexual Abuse and Exploitation Directive (CSAED).[206] The directive has set up minimum rules concerning the definition of criminal offences and sanctions in the area of child sexual exploitation and abuse.[207] It obliges "Member States to take the necessary measures to ensure the prompt removal of, or with appropriate safeguards block access to, web pages containing or disseminating child pornography. On that basis, Member States have implemented Notice-and-Takedown procedures through national hotlines, to which internet users can report child sexual abuse material that they find online."[208]

Shift of terminology

Two EP resolutions called to correct and replace the definition of child pornography with child sexual abuse material in order to reflect the broader scope that these crimes have.[209] Since then, these calls have been heard and the new definition was enshrined in the next legislation.

[203] Council framework Decision 2004/68/JHA of 22 December 2003 on combating the sexual exploitation of children and child pornography 2003 (OJ L).

[204] Council Decision of 29 May 2000 to combat child pornography on the internet 2000 (OJ L).

[205] Decision No 854/2005/EC of the European Parliament and of the Council of 11 May 2005 establishing a multiannual Community Programme on promoting safer use of the internet and new online technologies (Text with EEA relevance) 2005 (OJ L).

[206] Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA (n 189).

[207] Directorate-General for Internal Policies of the Union (European Parliament) and others, *Online Platforms' Moderation of Illegal Content Online: Laws, Practices and Options for Reform* (Publications Office of the European Union 2020) <https://data.europa.eu/doi/10.2861/831734> accessed 23 January 2023.

[208] de Streel and Husovec (n 92).

[209] 'European Parliament Resolution of 11 March 2015 on Child Sexual Abuse Online (2015/2564(RSP)' <https://www.europarl.europa.eu/doceo/document/TA-8-2015-0070_EN.html> accessed 23 January

The interim CSAM Regulation (2021)

Since the expansion of the notion of electronic communication services in the European Electronic Communication Code (EECC),[210] e-privacy now includes interpersonal communication services in its scopes such as WhatsApp, Instagram, and Messenger. The detection and reporting of CSAM by these services have clashed with the protection granted under the e-Privacy Dir.[211] To fix this issue, the EC has adopted an interim CSAM regulation in July 2021 which will last until August 2024.[212] This is the reason why the 2022 proposal has been released and is now in the EU policy-making pipeline to make sure to reach an agreement before the end of the interim text.

The EC proposal for the interim CSAM regulation[213] was criticised by the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS) as it did not contain enough safeguards against privacy threats and suffers from several legal gaps which include: "lack of clarity on the technologies that can be used, the proposal's failure to provide a legal basis for the processing, the uncertainty on the specificities of its measures, its wide scope and its failure to explicitly refer to transparency obligations."[214] Private companies get a considerable amount of responsibility without having enough transparency and accountability obligations to clarify the scope of these obligations. This creates risks of over compliance and false positives.

The EP adopted its position and brought clarifications in its amendments in relation to the technologies which can be used, the need to have a relevant GDPR legal basis for the processing

2023; European Parliament resolution of 14 December 2017 on the implementation of Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography (2015/2129(INI)) 2017.

[210] Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications Code (Recast) (Text with EEA relevance)Text with EEA relevance 2018.

[211] Charlotte Somers, 'The Proposed CSAM Regulation: Trampling Privacy in the Fight against Child Sexual Abuse?' (*CITIP blog*, 3 January 2023) <https://www.law.kuleuven.be/citip/blog/the-proposed-csam-regulation-trampling-privacy-in-the-fight-against-child-sexual-abuse/> accessed 20 January 2023.

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

[213] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a temporary derogation from certain provisions of Directive 2002/58/EC of the European Parliament and of the Council as regards the use of technologies by number-independent interpersonal communications service providers for the processing of personal and other data for the purpose of combatting child sexual abuse online 2020.

[214] Somers (n 211), EDPB-EDPS Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council Laying down Rules to Prevent and Combat Child Sexual Abuse | European Data Protection Board' <https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-042022-proposal_en> accessed 20 January 2023

of personal data, and the scope got clarified.[215] The text of the interim regulation was adopted in July 2021 and entered into force in August 2021. Some concerns remained even after the amendments, but the interim agreement was never challenged at the CJEU.[216]

The proposal for a new CSAM regulation (2022)

In May 2022, the EC released a proposal for a new regulation combating CSAM.[217] It builds on the 2011 Directive and the 2002 EU strategy for a more effective fight against child sexual abuse.[218] This comes at a time when COVID-19 has exacerbated the issue and reports show a dramatic increase in the reported cases of child sexual abuse.[219] This new proposal aims to replace the current system based on voluntary detection and reporting by companies. The proposal suggests imposing qualified obligations on providers of hosting services, interpersonal communication services, and other services concerning the detection, reporting, removing, and blocking of known and new online child sexual abuse material, as well as solicitation of children. This would solve the lack of harmonisation on rules and processes to detect CSAM content by the provider's services. In addition, the voluntary mechanism has proven inefficient to stop the spread as "with the vast majority of reports coming from a handful of providers, while a significant number take no action. Up to 95% of all reports of child sexual abuse received in 2020 came from one company, despite clear evidence that the problem does not only exist on one platform alone".[220]

[215] P9_TA(2021)0319 Use of technologies for the processing of data for the purpose of combating online child sexual abuse (temporary derogation from Directive 2002/58/EC) ***I European Parliament legislative resolution of 6 July 2021 on the proposal for a regulation of the European Parliament and of the Council on a temporary derogation from certain provisions of Directive 2002/58/EC of the European Parliament and of the Council as regards as the use of technologies by number-independent interpersonal communications service providers for the processing of personal and other data for the purpose of combatting child sexual abuse online (COM(2020)0568 — C9-0288/2020 — 2020/0259(COD)) P9_TC1-COD(2020)0259 Position of the European Parliament adopted at first reading on 6 July 2021 with a view to the adoption of Regulation (EU) 2021/… of the European Parliament and of the Council on a temporary derogation from certain provisions of Directive 2002/58/EC as regards the use of technologies by providers of number-independent interpersonal communications services for the processing of personal and other data for the purpose of combating online child sexual abuse 2021.

[216] Somers (n 211).

[217] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down rules to prevent and combat child sexual abuse 2022.

[218] European Commission, 'EU Strategy for a More Effective Fight against Child Sexual Abuse' <https://home-affairs.ec.europa.eu/policies/internal-security/child-sexual-abuse/eu-strategy-more-effective-fight-against-child-sexual-abuse_en> accessed 25 January 2023.

[219] 'Fighting child sexual abuse of children: Commission proposes new rules to protect children' (*European Commission*) <https://ec.europa.eu/commission/presscorner/detail/es/ip_20_2463> accessed 20 January 2023; Europol, 'Exploiting Isolation: Sexual Predators Increasingly Targeting Children during COVID Pandemic' (*Europol*) <https://www.europol.europa.eu/media-press/newsroom/news/exploiting-isolation-sexual-predators-increasingly-targeting-children-during-covid-pandemic> accessed 24 January 2023.

[220] 'Fighting child sexual abuse of children: Commission proposes new rules to protect children' (n 219).

The innovation of this proposal lies in the creation of a new obligation for service providers to detect, report, remove, block CSAM and alert the authorities. The proposal would now allow "court orders to require providers of end-to-end encrypted communication services, such as WhatsApp and Signal, to detect and report child pornography to law enforcement".[221] The providers will also have to assess and mitigate the risk of misuse of their service in a proportionate way. Risk assessments will be reviewed by Member States' authorities. The proposal follows a country of establishment principle where the authority of the country where the provider is established would get the competence of dealing with the case. "Companies having received a detection order will only be able to detect content using indicators of child sexual abuse verified and provided by the EU Centre".[222]

The proposal also creates a new independent EU Centre on Child Sexual Abuse. The Centre will be a hub of expertise providing information, and material to the online services providers. It will also collect best practices and help victims to take down CSAM content targeting them. The Centre will support the national law enforcement authorities and Europol in pre-analysing the providers' reports and channelling them promptly with the relevant authorities. The proposed regulation also aims to better protect, support and empower the victims. The Centre will set up an online support platform and it could for instance proactively search materials online and notify companies to take them down.

Since its release, the proposal for regulation was subject to feedback. The EC received 414 feedbacks and 81.6% of them come from EU citizens with a huge participation coming from Germany.[223] The text is now being debated and negotiated by EU policymakers (EP and Council).

- **Critical assessment**

The law has been described as "timely and historic, not just for Europe but for the world" by 90 child rights organisations in an open letter. It operates an important shift in the content moderation regulation of CSAM from a voluntary practice to binding obligations on providers. From a plural and scattered approach, the EU levels up its action and pushes for hard regulation. The creation of the EU Centre is also welcomed, especially as for now in the absence of such an

[221] Laura Kabelka, 'MEPs Sceptical on EU Proposal to Fight Online Child Sexual Abuse' (*www.euractiv.com*, 11 October 2022) <https://www.euractiv.com/section/digital/news/meps-sceptical-on-eu-proposal-to-fight-online-child-sexual-abuse/> accessed 25 January 2023.

[222] 'Controversial Proposal on Combating Child Sexual Abuse Online' <https://eucrim.eu/news/proposal-on-combating-child-sexual-abuse-online/> accessed 23 January 2023.

[223] European Commission, 'Feedback and Statistics: Proposal for a Regulation. Fighting Child Sexual Abuse: Detection, Removal and Reporting of Illegal Content Online' <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12726-Fighting-child-sexual-abuse-detection-removal-and-reporting-of-illegal-content-online/feedback_en?p_id=30786148> accessed 25 January 2023.

EU central organisation service providers can send their reports to, reports of abuse in the EU are sent to the United States (US) and then back to the EU law enforcement agencies.[224]

However well intended, the current 2022 proposal has been subject to criticisms from scholars, EU co-legislators, and civil society. The criticisms focus on risks that the proposal's provision brings to the proportionality principle, data protection, and the right to privacy. Rights are guaranteed at several levels of the European Union legal order. Firstly, the primary law level includes the Treaties of the European Union (TEU and TFEU) and the Charter of fundamental rights of the EU. Then, the secondary EU law includes the famous GDPR[225] and the E-privacy Directive.[226] As explained earlier, since the expansion of the notion of electronic communication services in the European Electronic Communication Code (EECC)[227], the e-privacy now includes interpersonal communication services in its scope such as WhatsApp, Instagram, and Messenger, which has created serious privacy concerns. It also raised concerns about overwhelming law enforcement authorities.

The proposal introduces a generalised scanning obligation for messaging services triggering mass surveillance practices and triggering an important debate in the EP. The EDPB-EDPS's opinion concluded that the Proposal raises serious concerns regarding the necessity and proportionality of the envisaged interferences and limitations to the protection of the fundamental rights to privacy and the protection of personal data.[228] The two institutions underlined that private companies enjoy a very broad margin of appreciation, which leads to legal uncertainty on how to balance the rights at stake in each case. This leaves too much room for potential abuse. In relation to the detection technologies in interpersonal communication services, the EDPB and EDPS also considered that they are disproportionate interferences due to the intrusiveness, probabilistic nature, and the error rates associated with such technologies. MEPs also underlined this aspect and pointed out that technology false positives could expose

[224] European Commission, 'EU Centre to Prevent and Combat Child Sexual Abuse. Why Oblige Platforms to Detect, Report and Remove Online Child Sexual Abuse' <https://home-affairs.ec.europa.eu/whats-new/campaigns/legislation-prevent-and-combat-child-sexual-abuse/eu-centre-prevent-and-combat-child-sexual-abuse_en> accessed 25 January 2023.

[225] Regulation (EU) 2016/679 of the European Parliament and of the Council - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation) 2016 [Regulation (EU) 2016/679] 88.

[226] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) 2009.

[227] Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications Code (Recast) (Text with EEA relevance) Text with EEA relevance.

[228] 'EDPB-EDPS Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council Laying down Rules to Prevent and Combat Child Sexual Abuse | European Data Protection Board' <https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-042022-proposal_en> accessed 20 January 2023.

innocent people to the screen of inspectors. Especially, in light of the mass of messages sent each day, the single smallest error rate could lead to the production of countless false reports.[229] In addition, data are missing or are too narrow to corroborate the 99% precision rate claimed by two companies providing data to the EC to enlighten the proposal drafting process.[230] Critics underline that independent tests should have been made. They say that only relying on the data of two companies' is not sufficient and wider tests should have been made and more data from various sources collected. The origin and quality of the data are indeed curious when a recent investigation by the Irish Council for Civil Liberties showed that the AI tools used to scan private communication to spot CSAM led to a low accuracy rate.[231] The AI tool had trouble identifying the context leading to false alarms cases where child peers above the age of sexual consent were sexting or simply knowing which legal age(s) of consent apply.[232]

Regarding detection obligations, they also conclude that measures permitting the public authorities to have access on a generalised basis to the content of a communication in order to detect grooming are more likely to affect the respect for private and family life and the protection of personal data. The EDPB and EDPS therefore suggest excluding grooming from the proposal. Furthermore, the scanning of audio communication was deemed particularly intrusive by the two data protection keepers and should be kept outside the scope of detection obligation.

In general, they believe that the proposal should operate a better balance between freedom of expression, right to privacy, and data protection and the fight against CSAM in order to meet societal needs such as having secure and private communication channels. They raise concerns that the provisions would endanger or weaken the use of encryption to protect the security of conversation. Edward Snowden expressed that encryption is a matter of life and death when it comes to whistleblowing, activism, and investigative journalism.[233] Furthermore, the collaboration and exchange of information between the new agency, Europol, and Data Protection Authorities (DPA) need to be further clarified.

[229] Laura Kabelka, 'EU Assessment of Child Abuse Detection Tools Based on Industry Data' (*www.euractiv.com*, 5 October 2022) <https://www.euractiv.com/section/digital/news/eu-assessment-of-child-abuse-detection-tools-based-on-industry-data/> accessed 25 January 2023.

[230] Laura Kabelka, (n 229).

[231] Olga Cronin, 'An Garda Síochána Unlawfully Retains Files on Innocent People Who It Has Already Cleared of Producing or Sharing of Child Sex Abuse Material' (Irish Council for Civil Liberties, 19 October 2022) <https://www.iccl.ie/news/an-garda-siochana-unlawfully-retains-files-on-innocent-people-who-it-has-already-cleared-of-producing-or-sharing-of-child-sex-abuse-material/> accessed 31 January 2023.

[232] Olga Cronin (n 231).

[233] Luca Bertuzzi, 'Whistleblowers Are Impossible without Encryption, Edward Snowden Says' (*www.euractiv.com*, 21 October 2021) <https://www.euractiv.com/section/data-protection/news/whistleblowers-are-impossible-without-encryption-edward-snowden-says/> accessed 25 January 2023.

In addition, concerns in relation to the country of establishment appeared. The goal would be to avoid one of the GDPR's biggest bottlenecks, where often only one single national authority would be responsible. The Council seemed to provide a solution in its proposed amendments with a procedure for cross-border removal orders.[234]

- **Future**

The future of CSAM regulation should make sure to empower and educate children and teenagers about CSAM and their rights in light of the new legislation. The literacy part is one of the cornerstones of the fight against this type of content, including educating law enforcement staff, society, and children/teenagers not to become a victim in the first place.[235] Another important recommendation is to clarify and be precise as much as possible when hard regulation is being adopted in order to ensure legal certainty and provide as much safeguard for this delicate subject as possible. This also joins the concerns in relation to the independence of authorities, human and financial resources, and relevant qualities that must be enshrined in the law. Another recommendation would be to conduct more comprehensive tests on the technologies and ensure having fair and diverse data, which can be screened and cross-checked by a wider panel of companies, child organisations, and researcher experts in the field. This will enable identifying the shortcoming of the technologies and their promises in a trustworthy manner. Some authors also worry about information that is not stored, such as Snapchat or the tool stories on Instagram.[236] The content is sent and received and auto-deleted after a few seconds, leaving no trace of the content.

In order to also fully address the CSAM challenges, policymakers need to think through the different distribution methods of CSAM to have regulation efficient for all. The P2P networks[237] are particularly attractive channels for perpetrators because they are free and publicly accessible. Furthermore, P2P networks do not use servers and can, therefore, transmit CSAM without oversight from electronic service providers.[238] The use of the Darknet increased with mobile devices, the rise of amateur content, social media, live streaming, newsgroups, and chat rooms. It is important to clarify and enshrine in the law whether the proposal covers cases of live-streaming formats. Especially, since this is a relatively new format and there is very scarce research regarding the detection of CSA on live streaming formats. It is crucial to ensure the various communication channels of CSAM are envisaged by the proposal. Research showed that

[234] Molly Killeen, 'EU Council Discusses Cross-Border Removal Orders to Fight Child Pornography' (*www.euractiv.com*, 21 November 2022) <https://www.euractiv.com/section/digital/news/eu-council-discusses-cross-border-removal-orders-to-fight-child-pornography/> accessed 25 January 2023.

[235] Europol, 'Exploiting Isolation: Sexual Predators Increasingly Targeting Children during COVID Pandemic' (n 219).

[236] Lee and others (n 201).

[237] Peer-to-Peer (P2P) networks are vast global systems for file sharing that are used by many people, and are typically known for acquiring music, movies, or other digital material from network users for free. Lee and others (n 201).

[238] Lee and others (n 201).

"CSAM detection applications were found to rely on image hash databases, keywords, web-crawler, detection based on filenames and metadata, and visual detection" and that "CSAM detection applications yielded the best results under multiple approaches combined, whereas deep-learning methods were demonstrated to outperform other ones for unknown CSAM." [239] Indeed, often, the techniques used are hash systems. It has proven to be efficient in P2P networks. However, it is quite easy to go around the hash detection systems.

Some also underline that policymakers should make sure not to tackle legal pornography or legitimate sexual material. Indeed, part of the biggest challenge of CSAM classification and detection is the presence of legal, pornographic material, as well as the presence of non-illegal material of children.

In addition, some processes should be envisaged to be elaborated in legislation to ensure collaboration between several departments due to the interdisciplinary nature of CSAM. This includes governments, law enforcement, researchers, tech companies, and organisations. Thus, safeguards must be elaborated in order to frame cautiously the scope, and methods of the collaboration.

To conclude, the privacy risks and personal data protection raised earlier must be well-balanced and safeguarded in the legislation. There are crucial aspects that cannot be overlooked despite the gravity of CSAM.

3.2.2.4 Hate speech

EU Code of Conduct on countering illegal hate speech online

- **Description of the main concepts**

In May 2016, the European Commission agreed with Facebook, Microsoft, Twitter and YouTube a "Code of conduct on countering illegal hate speech online"[240] and committed to fighting the dissemination of illegal hate speech. In the course of 2018, Instagram, Snapchat and Dailymotion joined the initiative, Jeuxvideo.com in January 2019, TikTok in 2020 and Linked in 2021. In May and June 2022, respectively, Rakuten Viber and Twitch announced their participation to the Code.[241] In particular, these intermediary service providers have made a series of commitments to:

- provide publicly available information on how to submit a notice flagging the hateful content;

[239] Lee and others (n 201).
[240] The Code of conduct on countering illegal hate speech online.
[241] https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

- put in place a clear and effective process to review notifications of "illegal hate speech" so they can remove or disable access to such content;
- review notifications on the basis of the Community Standards/Guidelines and the national transposition laws, and review the notifications within 24 hours;
- encourage the so-called 'trusted flaggers' system by providing training and support to the flaggers in order to ensure the quality of the notifications;
- strengthen communication and cooperation between the online platforms and the national authorities, and share best practices.[242,243]

- **Critical assessment**

The Code has faced massive criticism, especially from the freedom-of-expression and digital rights organisations – such as EDRi, ARTICLE 19, and the Center for Democracy & Technology, which warned that the Code could lead to more censorship by private companies, and, therefore, a chilling effect on freedom of expression.[244] Among many concerns raised, the following ones are the most fundamental. First, as mentioned above in the assessment of the e-Commerce Directive, the Code also imposes the 'de facto' regulatory role on online platforms. It puts companies – rather than the courts – in the position of having to decide the legality of content.[245] Moreover, as noted by Kuczerawy, by encouraging private companies to restrict speech of individuals the European Commission became an initiator of the interference with a fundamental right by private individuals.[246]

Content removals take place on the basis of definitions of 'hateful' or 'harmful' content, which are set forth by platforms themselves in their policies and Terms of Service. As pointed out by Bukovska, "hateful conduct" is a vague term that could encompass mere vulgar abuse.[247] Critics point out that platforms' understanding of these notions can go beyond, or even have no direct connection to the definitions established by the law.[248] In fact, platforms tend to base their policies on the most restrictive national law and apply them regardless of the jurisdiction, with the aim to minimise the risk of fines. This tendency could incentivise censorship and over-removal of content, with severe implications on the users' freedom of expression.

Second, lack of transparency in the reporting systems. The information provided by the platforms on their implementation of the Code of Conduct is incomplete, as it merely focuses

[242] Gellert and Wolters (n 90).
[243] EPRS, Polarisation and the use of technology in political campaigns and communication.
[244] Barbora Bukovská, 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online'.
[245] Bukovská (n 244).
[246] Aleksandra Kuczerawy, 'The Code of Conduct on Online Hate Speech: An Example of State Interference by Proxy?' (*CITIP blog*, 20 July 2016) <https://www.law.kuleuven.be/citip/blog/the-code-of-conduct-on-online-hate-speech-an-example-of-state-interference-by-proxy/> accessed 20 March 2023.
[247] Bukovská (n 244).
[248] EPRS, Polarisation and the use of technology in political campaigns and communication (n 243)

on the number and speed of removal, without actually explaining, for example, which percentage of the removed content was found 'illegal', and how much of it was later found to be the result of over-removal.

Third, there is a lack of sufficient safeguards against misuse of the notice procedure. There is also no procedure allowing to challenge wrongful removals. Lack of feedback on notifications reduces the users' understanding of what type of content is allowed or not online.[249] Moreover, the Code does not include any specific commitments to provide access to an appeal mechanism or other remedy for internet users whose content has been removed.

In that regard, a German judgement by the Federal Court of Justice is worth mentioning.[250] Facebook blocked a user's account and deleted comments because they violated the terms and conditions of the platform, which the user agreed to (prohibiting "hate speech"). The Court decided that the deletion of user contributions and account blocking in the event of violations of the communication standards set out in the terms are invalid. Facebook did not inform the user at least retrospectively about the removal of his content and about an intended blocking of his user account in advance. It also did not provide the reason for this. Simply put, while Facebook has the right to remove any posts and block user accounts that breaches the terms of service, it has to notify the user in question and allow for the opportunity to respond. The German case has illustrated the difficulty to strike a balance rights and interests in online content moderation. Facebook's freedom to conduct business and its own freedom of expression must be balanced with those of the users in such a way that the users' fundamental rights have the greatest possible effect, the Court said.[251] There must be objective reasons for the removal of content and the blocking of user accounts and "procedural protection of fundamental rights" including the clarification of the underlying facts behind the removal.[252]

Moreover, the Code provides that the content deemed as "illegal hate speech" should be taken down within 24 hours and there is no possibility for the user to contest the removal.[253] This leads to over-blocking practices.

It is worth mentioning that a short content removal deadline was subject to national courts decisions. In particular, the French Constitutional Court declared the so-called "Avia law" on

[249] EPRS, Polarisation and the use of technology in political campaigns and communication (n 243)
[250] Judgments of July 29, 2021 - III ZR 179/20 and III ZR 192/20
https://www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/2021/2021149.html
[251] 'Regulating Online Speech: Ze German Way' (*Lawfare*) <https://www.lawfareblog.com/regulating-online-speech-ze-german-way> accessed 20 March 2023.
[252] 'Regulating Online Speech: Ze German Way' (n 251).
[253] Bukovská (n 244).

hate speech, unconstitutional.[254] The law established two key obligations for the providers of online communication services: a) the obligation to remove within 1-hour terrorism or child-pornography content notified by the administrative authority; b) the obligation to remove hateful or sexual content flagged by users within 24 hours. The Constitutional Council that such a short deadlines and the high fines issued in case of non-compliance, encourages the removal of all content which has been flagged as potentially unlawful, even when this alleged unlawful content turns out to be perfectly lawful.[255] Such a mechanism constitutes a restriction to the right to freedom of expression and it is unconstitutional because of its unnecessity, inadequacy and lack of proportionality. The judgement provides an important lesson for the future regulations of 'hate speech' moderation online.

- **Future**

The European Commission and the IT companies should consider revising the Code of Conduct, in light of questions and implications for freedom of expression under the Code of Conduct. Thus, the companies should address these concerns, and they should be more transparent about their content moderation practices, including providing some case studies, i.e., qualitative analysis of their decisions and detailed information about the tools they use to moderate content, such as algorithms and trusted flagger schemes. The companies should also improve the internal complaints mechanisms, including those used for the wrongful removal of content or other restrictions on their users' freedom of expression. In general, individuals should be given detailed notice of a complaint and be provided with an opportunity for prompt redress. Internal appeal mechanisms should be clear and easy to find on company websites.[256] Some of these issues are tackled in the recent Digital Services Act (see Section 3.2.1.3).

Additionally, on 9 December 2021, the European Commission published an initiative to extend the list of EU crimes to hate speech and hate crime - whether because of race, religion, gender or sexuality, to establish minimum rules on the definition of criminal offences and sanctions in the areas of hate speech and hate crime.[257]

[254] Decision n° 2020-801 DC 18 June 2020 https://www.conseil-constitutionnel.fr/actualites/communique/decision-n-2020-801-dc-du-18-juin-2020-communique-de-presse

[255] Ilaria Buri, 'The Lesson of the French Constitutional Council on the Fight against Hate Speech and the Latest on the Upcoming Digital Services Act' (*CITIP blog*, 6 October 2020) <https://www.law.kuleuven.be/citip/blog/the-lesson-of-the-french-constitutional-council-on-the-fight-against-hate-speech-and-the-latest-on-the-upcoming-digital-services-act/> accessed 20 March 2023.

[256] Bukovská (n 244).

[257] Proposals to extend the list of EU crimes to all forms of hate crime and hate speech
In "A New Push for European Democracy"https://www.europarl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/file-hate-crimes-and-hate-speech

3.2.2.5 Disinformation

- **Description of the main concepts**

Perhaps, fake news and disinformation have always been there since the early days of the publishing industry and even before. However, the emergence of social media platforms and their widespread availability and accessibility made us more aware of their potential harms. Some scholars call this shift the 'post-truth' era,[258] and the term 'post-truth' even went on being declared the international word of the year in 2016.[259] Nevertheless, despite this overwhelming public and scholarly attention, disinformation is not a uniformly defined concept. There are many different definitions and scopes that have been coined so far.[260] Hence providing a legal definition of this polysemic term is not easy. For instance, the High-Level Expert Group on Fake News and Online Disinformation defines disinformation as "all forms of false, inaccurate, or misleading information designed, presented promoted to intentionally cause public harm or for profit." Accordingly, the definition "does not cover issues arising from the creation and dissemination online of illegal content (notably defamation, hate speech, and incitement to violence. Nor does it cover other forms of deliberate but not misleading distortion of facts such as satire and parody."[261] Accordingly, although there may be varying interpretations of the term by different actors, the consensus remains that not all content labelled as disinformation is necessarily illegal, but it can still be harmful. For this reason, some of the biggest tech companies (Facebook, Google, Twitter, Mozilla, and Microsoft) as well as the advertising industry, agreed on a Code of Practice on Disinformation in 2018. The Code is a soft law tool described as a voluntary, self-regulatory mechanism, with several commitments made by the signatories. Some of these commitments are directly related to content moderation practices, such as:

- Closing false accounts by developing clear policies regarding the identity and misuse of automated bots on their services;
- Investing in technologies to help internet users make informed decisions when receiving false information (e.g., reliability indicators/trust markers, reporting mechanisms);
- Prioritising relevant and authentic information; and
- Facilitating the finding of alternative content on issues of general interest.[262]

[258] Emiliana De Blasio and Donatella Selva, 'Who Is Responsible for Disinformation? European Approaches to Social Platforms' Accountability in the Post-Truth Era' (2021) 65 American Behavioral Scientist 825.

[259] 'Oxford Word of the Year 2016' (*Oxford Languages*), available at: <https://languages.oup.com/word-of-the-year/2016/> accessed February 20, 2023.

[260] Emine Ozge Yildirim, 'Silenced, Chilled, and Jailed: The New Turkish Law Criminalizes Disseminating 'Disinformation,' (VerfBlog, 20 October 2022), <https://verfassungsblog.de/silenced-chilled-and-jailed/>

[261] European Commission, 'A multi-dimensional approach to disinformation, Report of the independent High level Group on fake news and online disinformation,' (12 March 2018).

[262] Directorate-General for Internal Policies of the Union (European Parliament) and others (n 80).

In September 2020, the European Commission published its **assessment of the Code of Practice on Disinformation**.[263] Numerous positive impacts have been found, including platforms enforcing policies to prevent their services from being used to spread misrepresentative or misleading advertisements; reduced monetization incentives to disseminate disinformation online for economic gain, and an introduction of the label for sponsored political ads. However, **a number of shortcomings** have been identified such as a lack of key definitions, vague concepts, a narrow scope, combined with lack of enforcement and monitoring mechanisms which undermined the Code's impact and its potential for being a level playing field instrument.[264] First, the assessment points out the fragmented implementation and limited participation (only 16 signatories after almost two years of being in effect), lack of involvement of other relevant stakeholders, in particular from the advertising sector, and a regulatory asymmetry illustrated by the COVID-19 pandemic as two non-signatory platforms, Messenger and WhatsApp, were considered to be serious contributors of the spread of COVID-19 disinformation. Second, the absence of relevant key performance indicators (KPIs) to assess the effectiveness of platforms' policies to counter the phenomenon. A lack of commonly shared definitions and more precise commitments combined with a lack of enforcement and monitoring mechanisms undermine the Code's impact. The assessment also points out the lack of adequate complaint procedures and redress mechanisms for wrong content takedowns or account suspension following a presumed violation of signatories' disinformation policies and the lack of sufficient safeguards to ensure the protection of freedom of expression in practice. In short, the Code created a situation that encourages private entities to interfere with the freedom of expression of internet users, it therefore challenged the prohibiting the general monitoring of online content and questioned the EU acquis. It creates the incentives to restrict speech that might be critical or controversial but is not illegal under EU law. Importantly, the questions that could be raised would be: (i) who should decide what content is relevant, authentic, accurate, and authoritative? and (ii) who is responsible if the content is mislabelled?

In 2021, the EC issued a **Guidance for a revised Code of Practice on Disinformation**, which sought to address gaps and shortcomings and create a more transparent, safe, and trustworthy online environment. The Guidance also aimed at evolving the existing Code of Practice towards a co-regulatory instrument foreseen under the DSA. Following the Guidance, the updated version of the Code, the **strengthened Code of Practice on Disinformation**, had been signed and presented in 2022, with 34 signatories who have joined the revision process of the Code of 2018.

[263] European Commission, "Disinformation: EU Assesses the Code of Practice and Publishes Platform Reports on Coronavirus Related Disinformation", available at: <https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1568> accessed February 20, 2023

[264] Noémie Krack, 'Could Do Better! The European Commission's Assessment of the EU Code of Practice on Disinformation Is out.' (*KU Leuven Centre for IT and IP law*, 20 October 2020) <https://www.law.kuleuven.be/citip/blog/could-do-better-the-european-commissions-assessment-of-the-eu-code-of-practice-on-disinformation-is-out/> accessed 20 March 2023.

The strengthened Code contains 44 Commitments and 128 specific measures in the following areas:

- Cutting financial incentives for purveyors of disinformation;
- Broadening participation for a variety of diverse players with a role in mitigating the spread of disinformation;
- Ensuring transparency of political advertising;
- Ensuring the integrity of services;
- Empowering users;
- Empowering researchers;
- Empowering the fact-checking community;
- Transparency centre and Taskforce; and
- Strengthened Monitoring framework.

- **Critical assessment**

Until now the EU regulation efforts were quite cautious with a self-regulation approach.[265] The revised Code, "while stepping up the signatories effort to tackle disinformation and improving the measures, relies on the voluntary efforts of the signatories. This approach was chosen to safeguard freedom of expression, avoid over-regulation and censorship. This voluntary approach is nevertheless reinforced by the co-regulatory mechanism set up by the DSA."[266] Indeed, the new Code is closely tied to the newly adopted DSA.

The Code's preamble Para (i) clearly indicates that the Code aims to "become a Code of Conduct under Art. 35 of the DSA… regarding VLOPs that sign up to its commitments and measures." The preamble Para (j) adds that VLOPs that signed up to all commitments relevant and pertinent to "their services should be considered as a possible risk mitigation measure under Art. 27 of the DSA." According to Helberger and others, when these paragraphs are read with the DSA's preamble Para 68,[267] they trigger the question of whether the signatories could withdraw from the Code at any moment, considering the EC's endorsement of this voluntary instrument. Hence, it is uncertain if a signatory decides to withdraw from the Code, the EC would develop a negative judgement of that signatory in terms of compliance with the DSA.[268] Thierry Breton, the EU Commissioner for the internal market had underlined that very large platforms that repeatedly

[265] Noémie Krack, 'DSA Proposal and Disinformation - Should "Traditional Media" Be Exempted from Platform Content Moderation?' (KU Leuven Centre for IT and IP law, 7 December 2021) <https://www.law.kuleuven.be/citip/blog/dsa-proposal-and-disinformation-should-traditional-media-be-exempted-from-platform-content-moderation/> accessed 20 March 2023.

[266] Noémie Krack, 'MediaFutures Contributes in the Fight against Disinformation | Media Futures' (*MediaFutures*, 20 June 2022) <https://mediafutures.eu/mediafutures-contributes-in-the-fight-against-disinformation/> accessed 20 March 2023.

[267] Preamble Para. 68, the Digital Services Act.

[268] Natali Helberger and others, 'The EU's regulatory push against disinformation: What happens if platforms refuse to cooperate?,' (VerfBlog, 2022/8/05), <https://verfassungsblog.de/voluntary-disinfo/> accessed February 20, 2023.

break the Code and do not carry out risk mitigation measures properly risk fines of up to 6% of their global turnover.[269] But it remains unclear whether the EC would try to motivate or exert pressure on others to become signatories of the Code, even if they are initially reluctant to do so.

Besides this co-regulatory mechanisms, the DSA contains further provisions which will help combatting disinformation. This includes the modalities created for systemic risks assessment and mitigation (art. 34-35).[270] Then, crisis protocols (art. 48), some user empowerment measures (art.26 & 27) and increased transparency requirements (art.14, 15, 17,…) will also contribute to fight disinformation.[271]

Additionally, Commitment 39 of the Code states that the EC will be the responsible body for monitoring compliance with the Code. According to EU DisinfoLab, a non-profit organisation focused on tackling disinformation, it is not clear whether the EC has the capacity and necessary resources to do such monitoring effectively.[272] One should also be concerned about whether a political organisation, with differing interests and objectives, would be the right place to make such an assessment. Regardless, since the Code is not a binding instrument, it is hard to say that the assessment of compliance and deviation should be done by the courts. The practical application of the Code and the enforcement of the DSA are expected to provide further clarity in the upcoming years. For now, the most recent example is that in February 2023, the recently Elon Musk-acquired Twitter was warned by the EC that the platform's reporting falls short compared to the other signatories, "with no information on commitments to empower the fact-checking community."[273] The next batch of reports will be due in Summer 2023, as the signatories that signed up for the Commitments will need to provide further insight on the implementation and data covering the next 6 months following the initial report.

Furthermore, while there are some non-VLOP signatories to the Code, the amount is insufficient to mitigate the possible dissemination of disinformation by the non-VLOPs. The commitments also seem like they are targeting more VLOPs and their non-compliance. Moving forward, the EU legislator should consider the effects of disinformation spread not only by VLOPs but also by

[269] European Commission, 'Disinformation: Commission Welcomes the New Stronger and More Comprehensive Code of Practice on Disinformation' (*European Commission*, 16 June 2022) <https://ec.europa.eu/commission/presscorner/detail/en/IP_22_3664> accessed 20 March 2023
[270] ibid.
[271] ibid.
[272] 'Position of the EU DisinfoLab on the 2022 Code of Practice on Disinformation' (EU DisinfoLab, September 8, 2022), <https://www.disinfo.eu/advocacy/eu-disinfolabs-position-on-the-2022-code-of-practice-on-disinformation/> accessed February 20, 2023.
[273] European Commission, 'Code of Practice on Disinformation: New Transparency Centre Provides Insights and Data on Online Disinformation for the First Time,' <https://ec.europa.eu/commission/presscorner/detail/en/mex_23_723> accessed March 6, 2023.

other platforms. This hinges on risk analysis and providing measures and mechanisms encapsulating the level of risk, regardless of the status of the platform.

- **Future**

There are some terms in the Code that are not very clearly defined, or some concepts were left somehow vague. For example, according to the EU DisinfoLab, repeated use of the 'harmful disinformation' term raises questions on what is meant by such a definition. Therefore, the EC should clarify what harmful disinformation entails and what potential impact such harm caused by this type of disinformation could have.[274] The potential vagueness or lack of definition in prominent terms poses substantial uncertainty when preserving the right to freedom of expression while tackling online disinformation. Furthermore, even though the Code encourages non-VLOPs to subscribe to the commitments applicable to their services, the number of non-VLOP signatories is not enough considering the cost-benefit analysis of disinformation spread by them. The EC should recognize that VLOPs are not solely responsible actors in disinformation, and non-VLOPs' roles and responsibilities should also be clarified in an open manner, while also explaining their compliance and non-compliance mechanisms.

Additionally, as mentioned above, under the co-regulation approach, the strengthened Code is tied to the DSA in several ways. Therefore, the relationship between the Code, as a self-regulatory and volunteer instrument, and the DSA, a binding legislation, should be further clarified to avoid any uncertainty for the signatories on whether they could withdraw from the commitments they subscribed to or from the Code completely.[275] Lastly, the EC should not take over the job of monitoring compliance with the commitments, as their expertise may fall short. Instead, the Commission should look into establishing an impartial body, stripped of political interests and consisting of experts in this area equipped with adequate financial and otherwise relevant sources, that would oversee the compliance.

[274] EU DisinfoLab (n 272).

[275] Natali Helberger and others (n 268).

4 Alternative approaches and future trends in content moderation

Content moderation is a multifaceted coin; it is not only about the legislation and framework initiatives of States, European, or International institutions. As a matter of fact, given the mass of content and the boom of user-generated content, intermediary service providers came up with some of their own ways to address content moderation challenges. This inspired further competitors and actors active in the same market, potentially leading the way towards cascade content moderation initiatives or mechanisms. In this section, we will analyse how end-user and community moderation are being used and how the Facebook Oversight board works. In addition, the model of the Social Media Council will also be analysed. This section aims to provide a better understanding of these content moderation initiatives, their opportunities, and shortcomings.

4.1 End-user moderation or Community-led moderation

4.1.1 Self-moderated communities

4.1.1.1 Wikipedia

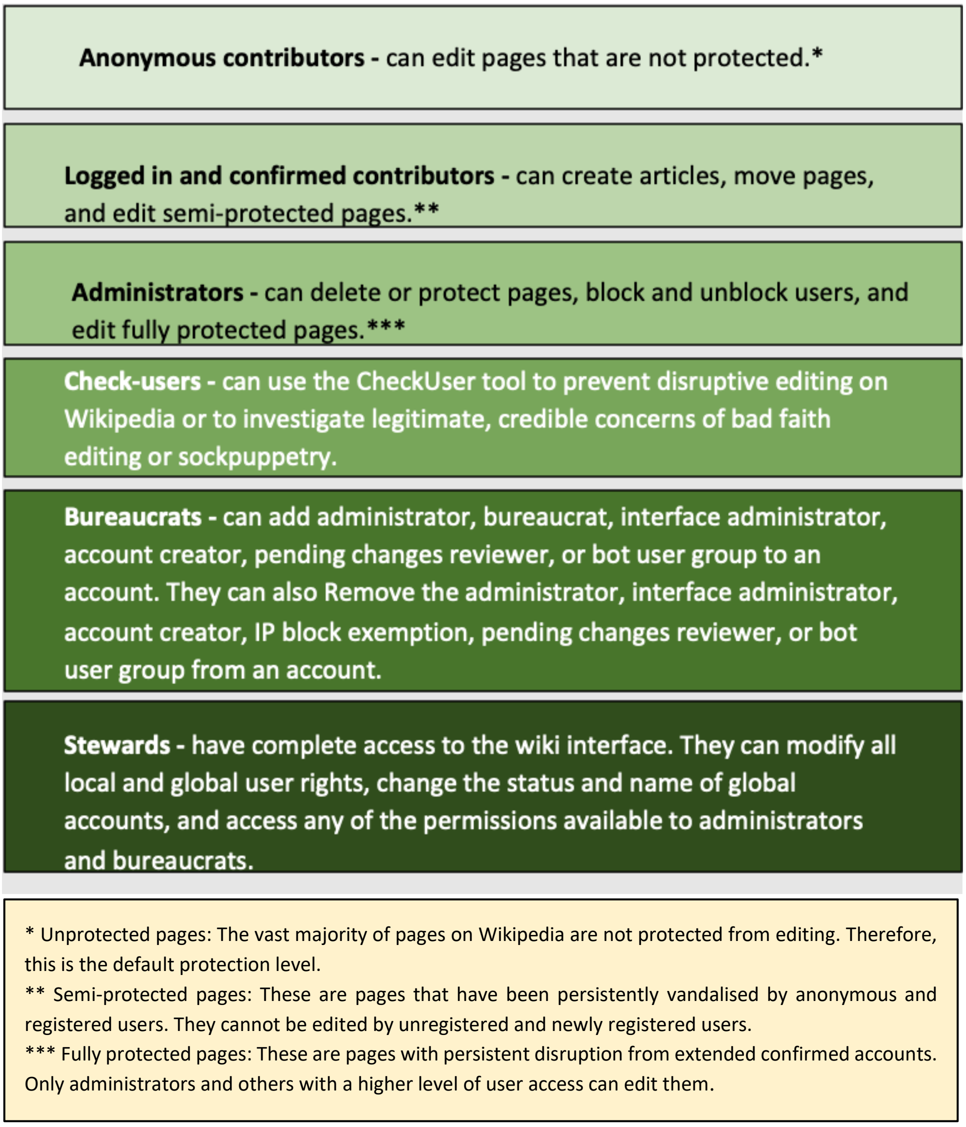
As the largest free online encyclopaedia in the World, Wikipedia, one of the many sister projects supported by the Wikimedia Foundation (WMF), is a volunteer community moderated platform. At the moment, Wikipedia has over 300 language versions, of which the English, Cebuano, German, Swedish, and French Wikipedias are the largest active communities.[276] According to the WMF, anyone can edit and improve articles on Wikipedia, as long as the content is written from "a neutral point of view and attributed to a reliable source."[277] WMF merely hosts content and is not involved in editing or creating any content. Thus, content moderation on Wikipedia hinges on volunteers consisting of administrators and editors, along with bots and monitoring tools. It is also important to note that the rules of content moderation on Wikipedia could also differ in different language versions. Therefore, this part will focus on the rules concerning the English version of Wikipedia as a reference.

As a self-moderated platform, Wikipedia does not have moderators or automated content recognition tools governed by the platform itself. On the human level, there are volunteer editors who could write and edit Wikipedia pages. There are some editor categories where they are granted a higher level of user access to moderation of the platform. For example,

[276] Ben Wagner and others, 'Reimagining Content Moderation and Safeguarding Fundamental Rights' (May 2021), <https://www.greens-efa.eu/files/assets/docs/alternative_content_web.pdf> accessed February 22, 2023.

[277] 'Our Work' (*Wikimedia Foundation,* January 19, 2023) <https://wikimediafoundation.org/our-work/> accessed March 6, 2023.

administrators, also known as system operators (sysops), are given access to restricted technical features, such as protecting and deleting pages and blocking other editors.[278] More information on this can be found in Figure 4 below.



Anonymous contributors - can edit pages that are not protected.*

Logged in and confirmed contributors - can create articles, move pages, and edit semi-protected pages.**

Administrators - can delete or protect pages, block and unblock users, and edit fully protected pages.***

Check-users - can use the CheckUser tool to prevent disruptive editing on Wikipedia or to investigate legitimate, credible concerns of bad faith editing or sockpuppetry.

Bureaucrats - can add administrator, bureaucrat, interface administrator, account creator, pending changes reviewer, or bot user group to an account. They can also Remove the administrator, interface administrator, account creator, IP block exemption, pending changes reviewer, or bot user group from an account.

Stewards - have complete access to the wiki interface. They can modify all local and global user rights, change the status and name of global accounts, and access any of the permissions available to administrators and bureaucrats.

* Unprotected pages: The vast majority of pages on Wikipedia are not protected from editing. Therefore, this is the default protection level.
** Semi-protected pages: These are pages that have been persistently vandalised by anonymous and registered users. They cannot be edited by unregistered and newly registered users.
*** Fully protected pages: These are pages with persistent disruption from extended confirmed accounts. Only administrators and others with a higher level of user access can edit them.

Figure 4: Overview of the categories of contribution status that can be awarded to Wikipedia users[279]

When it comes to the use of automated tools and bots, Wikipedia follows more of a decentralised software approach. The MediaWiki software, which is the software platform

[278] 'Administration' (*Wikipedia,* November 24, 2022) <https://en.wikipedia.org/wiki/Wikipedia:Administration> accessed March 6, 2023

[279] 'Protection Policy' (Wikipedia, February 24, 2023) <https://en.wikipedia.org/wiki/Wikipedia:Protection_policy> accessed March 6, 2023.

Wikimedia projects are built on, allows for the volunteer community to develop third-party tools, scripts, bots, and other external software.[280] This does not mean there are no centralised governance processes on Wikipedia, as fully automated bots could be utilised for large-scale editing. However, even with these processes, bot approval decisions are made by the volunteer-based local governance mechanisms of the respective language version of Wikipedia. Therefore, apart from providing the infrastructure, the WMF staff has no role in this type of editing or content creation as well.[281] For instance, ORES is a website and application programming interface (API) that provides machine learning as a service for Wikipedia and other Wikimedia projects developed by the WMF Staff.[282] The system assists human editors and automates some critical, time-consuming tasks like "detecting vandalism and removing edits made in bad faith."[283] Accordingly, with the assistance of technological tools, human editors are able to evaluate whether the edit or content created would require any intervention,[284] while the system continuously learns to distinguish between good faith and bad faith edits.[285]

Additionally, the WMF states that contributors are legally responsible for all contributions and edits under the laws of the US and other applicable laws, which may include the laws where the contributors live or where they view or edit content.[286] While most content, including shocking and offensive, is allowed on Wikipedia under the WMF's mission, defamation, harassment, threatening, and copyright-infringing content is prohibited on the platform.[287]

It is also important to note that being a self-moderated community platform, with a combination of algorithmic and human moderation, does not rule out the issues arising from content moderation in general. While Wikipedia follows a very different approach compared to other big platforms, content hosted by Wikipedia could still be biased and inaccurate, and in rare cases, it might take a while for malicious, defamatory, or disinformation content to be removed by the volunteer community.[288]

[280] Aaron Halfaker and R. Stuart Geiger, 'Ores: Lowering Barriers with Participatory Machine Learning in Wikipedia'(2020) 4 Proceedings of the ACM on Human-Computer Interaction 1.

[281] ibid.

[282] 'Ores - MediaWiki' (*Powered by MediaWiki,* June 8, 2022) <https://www.mediawiki.org/wiki/ORES> accessed February 20, 2023.

[283] 'Non-Free Content' (*Wikipedia,* January 15, 2023) <https://en.wikipedia.org/wiki/Wikipedia:Non-free_content> accessed February 24, 2023.

[284] Ben Wagner and others (n 276).

[285] Paul B. de Laat, 'The Use of Software Tools and Autonomous Bots against Vandalism: Eroding Wikipedia's Moral Order?' (2015) 17 Ethics and Information Technology 175.

[286] 'Terms of Use' (*Wikimedia Foundation Governance Wiki*) <https://foundation.wikimedia.org/wiki/Terms_of_Use/en> accessed March 6, 2023.

[287] 'What Wikipedia Is Not' (*Wikipedia,* March 4, 2023) <https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#Wikipedia_is_not_censored> accessed March 6, 2023.

[288] Yumiko Sato, 'Non-English Wikipedia is rife with misinformation. Here's how to fix it' (Fast Company, 27 July 2021) <https://www.fastcompany.com/90666412/non-english-wikipedia-misinformation accessed> accessed March 17, 2023. Also see, 'Who Gets To Be 'Notable' — And Who Doesn't? Gender

Wikipedia and Copyright-protected content

The Wikimedia movement is built upon the mission of free culture and open knowledge. Therefore, content on Wikipedia could be used, reused, and modified freely without the prior permission of the authors of such content. This, however, does not mean that content on Wikipedia can infringe copyright laws. Text and non-text materials should be compatible with the open license requirements of the platform. Such content could be in the public domain or be licensed under Creative Commons (CC BY-SA) or GNU Free Documentation Licenses.[289] In some circumstances, non-free content could be used on English Wikipedia without first acquiring permission from the copyright holder strictly within the framework of the US legal doctrine of fair use. However, the fair use doctrine is inherently a US law concept, recognized in only a few other jurisdictions, that does not find the same applicability in EU law and most member states' legal framework.

Furthermore, DMCA §512 mentioned in Section 3.2.2.2 is applicable to Wikipedia for content removal requests, whereas Art. 17 of the CDSM is not deemed applicable due to the exceptions listed in Recital 62, which excludes not-for-profit online encyclopaedias from the scope of OCSSPs.

Wikipedia vs. the DSA

Regarding whether the DSA is applicable to Wikipedia, there are a few aspects that need to be considered. Between Aug. 2022 and Jan. 2023, Wikipedia had around 150 million monthly active recipients of the service in the EU region.[290] The number is well above the 45 million threshold. Therefore, it is possible that it will be designated as a VLOP and will be overseen by the EC, instead of the Digital Services Coordinators of member states. Under Art. 2 of the DSA, while community voluntary moderation is not explicitly recognised, it is implied that the DSA will mainly focus on moderation done by the service providers.[291] The obligation of providing clear and understandable terms of services and enforcing them accordingly, imposed on the service providers under Art. 12, will also not be applicable to the editing community of Wikipedia and will not prevent them from moderating the platform. Therefore, the DSA will not interfere with community content moderation.

Wikipedia will still need to comply with some obligations set forth in the DSA, such as performing a regular assessment of systemic risks concerning content like disinformation and illegal content

Bias On Wiki' (National Public Radio, 13 July 2021) <https://www.npr.org/2021/07/13/1015754856/who-gets-to-be-notable-and-who-doesnt-gender-bias-on-wiki> accessed March 17, 2023.

[289] 'Copyrights' (*Wikipedia,* December 20, 2022) <https://en.wikipedia.org/wiki/Wikipedia:Copyrights> accessed February 24, 2023.

[290] 'EU DSA USERBASE Statistics' (*Wikimedia Foundation Governance Wiki*) <https://foundation.wikimedia.org/wiki/Legal:EU_DSA_Userbase_Statistics> accessed March 6, 2023.

[291] Dimi Dimitrov, 'DSA: Political Deal Done' (*Free Knowledge Advocacy Group EU*April 26, 2022) <https://wikimedia.brussels/dsa-political-deal-done/> accessed February 25, 2023.

and putting in place relevant mitigation measures for identified risks subject to independent audits.[292] However, as Wikipedia does not run ads, other obligations concerning advertising will not be applicable to Wikipedia. Finally, while the EC could impose a fee on VLOPs, not-for-profit organisations like WMF are exempt from such a fee.[293]

4.1.1.2 Discord

Discord is a Voice Over Internet Protocol and instant messaging social platform, with features like servers (Discord communities), channels, private messaging, and video call/streaming. Discord is also a community content moderation platform, relying on the admins of servers to handle moderation. The size of Discord servers ranges from small groups to massive communities with thousands of users.[294] Server creators could create hierarchical roles granting different permissions to these roles, including removing or muting users and banning users in some cases.[295] While third-party bots have been utilised by Discord server moderators and admins for a while now, Discord has recently introduced a moderation tool called ’AutoMod,’ to assist admins and moderators keep their servers safe. Community content moderation is a time-consuming task involving the fear of not knowing what would happen if one took a break from the server for a while. With this tool, Discord promises that keyword filters, pre-set provided by Discord and custom tailored by the moderation team of the server, will detect harmful content before it is even posted on the server or channel.[296] Therefore, it provides flexibility for the moderation team of the server to be able to handle this content when they are back while keeping the server ’safe’ in the meantime.

While Discord’s community moderation model looks promising, it does not come without problems. For instance, Discord’s content came under scrutiny when it was realized that extremist users and groups, such as Alt-right and neo-Nazis, had utilised the platform to spread harmful content and organize the white supremacist Charlottesville attack.[297] After this incident, Discord banned servers promoting Nazi and white supremacist ideologies.[298] Another example

[292] ibid.

[293] ibid.

[294] Jialun Aaron Jiang and others, ‘Moderation Challenges in Voice-Based Online Communities on Discord’ (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1.

[295] ibid.

[296] ‘Meet Your Newest Community Moderator: AutoMod Is Here’(*Discord Blog,* October 31, 2022) <https://discord.com/blog/automod-launch-automatic-community-moderation> accessed February 25, 2023.

[297] Kevin Roose, ‘This Was the Alt-Right's Favorite Chat App. Then Came Charlottesville’ (*The New York Times,* August 15, 2017) <https://www.nytimes.com/2017/08/15/technology/discord-chat-app-alt-right.html> accessed February 25, 2023.

[298] Megan Farokhmanesh, ‘White Supremacists Who Used Discord to Plan Charlottesville Rally May Soon Lose Their Anonymity’ (*The Verge,* August 7, 2018) <https://www.theverge.com/2018/8/7/17660308/white-supremacists-charlottesville-rally-discord-plan> accessed February 28, 2023.

would be banning the pro-Donald Trump server two days after the U.S. Capitol attack due to its overt connection to an online platform used to incite violence, plan an armed insurrection in the US, and spread harmful misinformation related to 2020 U.S. election fraud,[299] despite not having enough evidence that the server was involved with the riot at all. However, it was found out that after Reddit's ban of the r/TheDonald subreddit months before the election,[300] former redditors had utilised Discord.[301] There are several other examples that could make one question whether Discord's community moderation model works efficiently, but content moderation, whether centralised or decentralised, is a hard task requiring taking into consideration multiple dimensions. Thus, Discord is not the only platform struggling in the face of a massive shift regarding content sharing.

Discord vs. the DSA

According to the recent data from Discord, the average number of monthly active recipients of the platform in the EU between July 2022 and December 2022 is well below 45 million.[302] Though this amount is lower than the threshold to be classified as a VLOP under the DSA, the platform and its services will still be held to certain obligations under the DSA. Their obligations, the extent of those obligations, and whether they qualify for the definition of VLOPs will be clearer after the EC designation and when the DSA becomes applicable.

Similar to the case with Wikipedia, the DSA should not interfere with Discord's community content moderation. However, if the platform qualifies as a VLOP, it will still need to comply with the requirements like regular systemic risk assessment and incorporate relevant mitigation measures for identified risks. Additionally, at the moment, Discord does not seem to be running advertisements. This means that the platform will not need to be bound by obligations set forth concerning advertising.

[299] Jay Peters, 'Discord Bans pro-Trump Server 'the Donald'' (*The Verge,* January 9, 2021) <https://www.theverge.com/2021/1/8/22221579/discord-bans-the-donald-server-reddit-subreddit> accessed February 26, 2023.

[300] Also *see*, "The popular message-board website Reddit, for example, grants substantial autonomy to its various subreddits, each of which has its own moderators. Indeed, Reddit is frequently held up as the most prominent example of bottom-up, community-based content moderation. Because Reddit can moderate any piece of content indeed, to ban a subreddit outright - no matter whether the subreddit moderator agrees, it is subject to public pressure to do so. Perhaps the most famous example is Reddit's banning of the controversial pro-Trump r/The_Donald subreddit several months before the 2020 election." Alan Z. Rozenshtein, 'Moderating the Fediverse: Content Moderation on Distributed Social Media," SSRN'<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213674> accessed 24 January 2023.

[301] Jay Peters (n 299).

[302] Discord Trust & Safety Team, 'Digital Services Act - Information on Average Monthly Active Recipients in the European Union,' <https://support.discord.com/hc/en-us/articles/12477677109143-Digital-Services-Act-Information-on-Average-Monthly-Active-Recipients-in-the-European-Union> accessed February 28, 2023.

Interim Conclusion on Self-moderated Communities

While in decentralised content moderation communities, users have higher confidence in distributed moderation over centralised moderation due to the reason that they are closer to the moderator,[303] there is also a set of challenges community moderation brings. Firstly, as the voluntary editors and moderators usually lack relevant expertise, along with their possible personal biases, they may be "incapable of making decisions representative of the community ideal."[304] Whereas traditional centralised moderation is considered to be more consistent, as moderation is mainly conducted by experts. Additionally, decentralised moderation approach may leave minorities in a disadvantageous situation, as decision-making could disproportionately favour majority norms.[305] Furthermore, according to some studies, it is shown that community moderation consists of a burdensome workload to some extent, with community moderators needing to spend a substantial amount of time.[306] As a result, community moderation could pose a real challenge as it constitutes a trade-off between improving efficiency and having key expertise on the topic.

4.1.2 Content moderation in fediverse

The term "fediverse", a portmanteau of "federation" and "universe", refers collectively to the protocols, servers, and applications that enable decentralised social media.[307] The topic of fediverse and content moderation from the legal point of view is currently under-researched. Among a few academic analyses, Alan Z. Rozenshtein provides the following overview on the subject. He points out that the most important feature of fediverse's protocol, namely ActivityPub - which powers the most popular fediverse apps - is that it is decentralised.[308] The servers - generally called "instances" - used to send content around the network are independently owned and operated. Anyone can create and run an instance as long as they follow the ActivityPub protocol. ActivityPub's decentralised nature means that an instance can choose what content flows across its network. Consequently, each instance can use different content-moderation rules and standards.[309] An instance can choose to block certain users, types of content (e.g., videos or images), or entire instances. But no instance can control the behavior

[303] Joseph Seering and others, 'Moderator Engagement and Community Development in the Age of Algorithms' (2019) 21 New Media & Society 1417.

[304] Jialun 'Aaron' Jiang (n 32); Sarah A Gilbert, ''I Run the World's Largest Historical Outreach Project and It's on a Cesspool of a Website.' Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians' (2020) 4 Proceedings of the ACM on Human-Computer Interaction 19:1.

[305] Stefanie Duguay and others, 'Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine' (2020) 26 Convergence 237.

[306] Eshwar Chandrasekharan and others, 'The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales' (2018) 2 Proceedings of the ACM on Human-Computer Interaction 32:1.

[307] Alan Z. Rozenshtein 'Moderating the Fediverse: Content Moderation on Distributed Social Media', SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213674> accessed 24 January 2023.

[308] Alan Z. Rozenshtein (n 307).

[309] Alan Z. Rozenshtein (n 307).

of any other instance, and there is no central authority that can decide which instances are valid or that can ban a user or a piece of content from the ActivityPub network entirely. Rozenshtein calls this a model of *content-moderation subsidiarity*. A key guarantor of such a model is the ability of users to switch instances if, for example, they are dissatisfied with how their current instance moderates content. When a user decides to move instances, they migrate their account data - including their blocked, muted, and follower user lists and post history - and their followers will automatically refollow them at their new account.[310]

This section uses the Mastodon project as a case study for content moderation in fediverse. Mastodon is the largest federating social network.

4.1.2.1 Mastodon

Although the organization that runs the Mastodon project requires each instance to follow certain high-level content moderation guidelines,[311] each Mastodon instance chooses its own content moderation policies. The rules governing these policies vary greatly: some may be far more restrictive than those of the VLOPs, and some less. As already explained, content moderation subsidiarity means that if a user disagrees with the rules of a chosen Mastodon instance, they can easily switch to another instance with other content moderation rules. This is in contrast to how VLOPs operate: centralised platforms, 'by their nature'[312] must decide on a single content moderation standard, which different users may find either under- or overinclusive, leading to mass dissatisfaction.

Mastodon vs. the DSA

The question raises how to classify Mastodon under the DSA. Is Mastodon considered to be a single service, or are instances essentially service providers that just happen to use a shared protocol? Husovec argues that given that the protocol is open, and Mastodon does not control who can create an instance, only instances (servers) should be regarded as services.[313] Mastodon instances will, therefore, be likely considered a 'hosting service'. Thus, each of these instances will need to comply with a set of minimum obligations for intermediary and hosting services, including having a single point of contact and legal representative, providing clear terms and conditions, publishing bi-annual transparency reports, having a notice and action mechanism

[310] Alan Z. Rozenshtein (n 307).

[311] The "Mastodon Server Covenant," provides: "All Mastodon servers we link to from our server picker commit to the following: [...] active moderation against racism, sexism, homophobia and transphobia. Users must have the confidence that they are joining a safe space, free from white supremacy, anti-semitism and transphobia of other platforms." 'Mastodon Server Covenant for Joinmastodon.Org' <https://joinmastodon.org/covenant> accessed 26 January 2023.

[312] Konstantinos Komaitis, 'Can Mastodon Survive Europe's Digital Services Act?' (*Tech Policy Press*, 16 November 2022) <https://techpolicy.press/can-mastodon-survive-europes-digital-services-act/> accessed 26 January 2023.

[313] Eli Cohen Lawson, 'New Research Shows Metaverse Is Not Safe for Kids' (*Center for Countering Digital Hate | CCDH*, 30 December 2021) <https://counterhate.com/blog/new-research-shows-metaverse-is-not-safe-for-kids/> accessed 26 January 2023.

and, communicating information about removals or restrictions to both notice and content providers. Should they exceed a threshold of micro and small enterprises, they may be regarded as an ‘online platform’, subject to more stringent DSA obligations. If any single instance reaches 45 million monthly active users, then it can become a VLOP. That instance would need to proceed to the implementation of additional requirements, including a complaint handling system, cooperation with trusted flaggers and out-of-court dispute bodies, enhanced transparency reporting and the adoption of child protection measures, as well as the banning of dark patterns.

Despite many advantages, fediverse is not a panacea to all content moderation problems. Because there is no centralised fediverse authority, there is no way to fully exclude even the most harmful content from the network. Moreover, fediverse administrators will generally have fewer resources, as content moderation is a voluntary-run type of service. Much will therefore depend on whether and how the decentralised content moderation framework scales.

4.1.3 Content moderation in the metaverse

Although there is no official definition, the metaverse can be described as “an immersive and constant virtual 3D world where people interact by means of an avatar to carry out a wide range of activities.”[314] Facebook’s VR Metaverse is just one example of such metaverse world. With great opportunities in the metaverse come great risks. The nature of the metaverse poses many challenges when it comes to addressing liabilities, combating illegal and harmful practices and misleading advertising practices, and protecting intellectual property rights.[315]

In the context of this report, the most pressing are the challenges which augmented and virtual reality will create for content moderation. Those include questions on how to tackle verbal harassment or hate speech in a virtual space, inappropriate actions from avatars that simulate sexual harassment or assault, pornographic content modelled on avatars, or misinformation or defamatory content generated using augmented reality. Research conducted by the Center for Countering Digital Hate (CCDH) shows that content on the VR Chat -part of Facebook’s VR Metaverse -contains abuse, harassment, racism and pornographic content.[316] CCDH researchers found that users, including minors, are exposed to abusive behavior every seven minutes. Such behaviour included: minors being exposed to graphic sexual content; bullying, sexual harassment and abuse of other users; minors being groomed to repeat racist slurs and extremist talking points; threats of violence. Researchers identified 100 potential violations of Facebook’s policies for VR in 11 hours and 30 minutes of recordings of user behavior in the app.[317] Moreover,

[314] EPRS, Metaverse. Opportunities, risks and policy implications

[315] EPRS (n 314).

[316] Lawson (n 313).

[317] Lawson (n 313).

cases of women being sexually harassed on Meta's VR social media platform have already been documented.[318]

Researchers show that the metaverse is even more difficult to moderate than social media platforms because it takes the existing content moderation problems and amplifies them even further.[319] In the metaverse, the users use voice chat and gesture, not text that will exist for long periods. Such content is harder to filter. As the EP study suggests, new approaches and technologies will need to be developed if future VR worlds are to be safe spaces for players.[320]

We can assume that some platforms will take a top-down approach to content moderation. This will require the massive-scale use of automated systems. However, as explained in Section 3.1.3, there are technical limitations of these tools which risk rendering content moderation in the metaverse ineffective. The most serious risk is perhaps the lack of understanding of the context. Slight behavioral changes or the use of symbols that exploit the algorithms' lack of comprehension of context.[321] "Algospeak" which refers to code words or turns of phrase is becoming an increasingly common phenomena across the internet as people seek to bypass content moderation tools on social media platforms such as TikTok, YouTube, Instagram and Twitch.[322] For instance, in many online videos, it's common to say "unalive" rather than "dead," "SA" instead of "sexual assault".[323] Moreover, relying on AI to monitor what people say and do in the metaverse would require every second of every interaction to be monitored and analysed. Similar problems have already arisen in the context of live audio content moderation on services such as Clubhouse or Twitter Spaces. Moreover, this raises privacy concerns and would require massive amounts of computing power.[324]

Other platforms may choose to adopt more of a decentralised approach that allows communities and volunteers to moderate the content. As shown above, this approach has seen

[318] 'Tanya Basu The Metaverse Has a Groping Problem Already | MIT Technology Review' <https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/> accessed 7 March 2023.

[319] Ryan Hsu 'Meet the New 'verse, Same as the Old 'verse: Moderating the "Metaverse"' (*Georgetown Law Technology Review*, 2 May 2022) <https://georgetownlawtechreview.org/meet-the-new-verse-same-as-the-old-verse-moderating-the-metaverse/GLTR-05-2022/> accessed 26 January 2023.

[320] EPRS (n 314).

[321] EPRS (n 314).

[322] Taylor Lorenz "Algospeak" Is Changing Our Language in Real Time - The Washington Post' <https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/> accessed 21 February 2023.

[323] Taylor Lorenz (n 323).

[324] Ryan Hsu (n 319).

some success in platforms like Wikipedia or Reddit. However, community-led moderation can lead to a lack of platform-wide standards and human moderators' burnouts.[325]

Metaverse vs. the DSA

The question raises whether the newly adopted EU content moderation rules in the DSA would apply to illegal or harmful metaverse content. As provided by the EP, the need to further amend EU law cannot be ruled out, since the topic of virtual reality is not specifically addressed in the DSA.[326] However, some point out that there is no question that the DSA does apply to games.[327] By extensions, other virtual 3D worlds are also covered by the DSA, at least as an "intermediary service", often also as an "online platform". However, it is desirable to define more clearly the extent to which virtual 3D worlds fall within the scope of the DSA.[328]

Additionally, it should also be mentioned that the AI Act proposal prohibits placing on the market, putting into service or use an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm. In this provision, the notion of 'subliminal techniques' is not defined, which makes the scope of application of this provision far from clear. One may wonder whether and to which extent metaverse practices fall within the scope of this provision. Additionally, the provision requires a person's behavior to be "materially distorted". It is ambiguous what this concept would mean in the context of the virtual world.

4.2 Accountability initiatives

Self-regulation initiatives seem to flourish in the content moderation landscape and provide some interesting concepts to study. Between the PR move and a real ambition to address some part of the content moderation challenges by themselves, scholars are investigating whether these initiatives would be beneficial or superficial for content moderation regulation. Among them, authors investigated whether self-regulation was beneficial and provided some general guidelines for when and how specific types of platform businesses could rely on self-regulation to address challenges more effectively.[329] In the following sub-section, we will investigate the promises and disadvantages of the Facebook Oversight Board and the concept of social media

[325] Juan Londoño 'Lessons from Social Media for Creating a Safe Metaverse' <https://itif.org/publications/2022/04/28/lessons-social-media-creating-safe-metaverse/> accessed 26 January 2023.

[326] EPRS (n 314).

[327] Julian Jaursch 'Opinion Piece: The DSA Also Works "in the Metaverse" – If It Is Enforced Well' (14 December 2022) <https://www.stiftung-nv.de/en/publication/opinion-piece-dsa-also-works-metaverse-if-it-enforced-well> accessed 26 January 2023.

[328] Julian Jaursch (n 327).

[329] Michael A Cusumano, Annabelle Gawer and David B Yoffie, 'Can Self-Regulation Save Digital Platforms?' (2021) 30 Industrial and Corporate Change 1259.

council developed by Article 19, an international human rights organisation that works to defend and promote freedom of expression and freedom of information worldwide.

4.2.1 Facebook Oversight Board

In 2018, in order to improve its content moderation decision, Meta (at the time Facebook) announced the establishment of the Oversight Board (OB). The OB can be categorised as a platform of self-governance for content moderation and hence being considered as self-regulation. The board is governed by two main documents, namely the OB Charter having primacy over the OB Bylaws.[330] The initial formation of the Board is composed of co-chairs selected by Facebook which then jointly with Facebook have selected candidates for the board (article 8 of the Charter). The OB trustees will appoint the members. Members of the public can recommend candidates for the board, but they must be formally appointed by the trustees.

The institution's mandate is plural but focuses mainly on the review and the issuance of binding decisions on content moderation decisions coming from Facebook and Instagram to remove or uphold content. The OB is not the extension of the Meta content review process. Its review is reserved for a selection of highly emblematic cases and determines if decisions were made in accordance with Meta's stated values and policies. Only few cases are actually taken and reviewed by the board.[331] In addition, the OB can issue non-binding recommendations about the platform's policies.

Practically, for a case to appear on the OB's desk both the users, and Meta must first agree about the review. Then, among those cases, it must be selected by the Board itself. Once the case is selected, a jury of five panellists is being set up. Different information is considered by the panel for reaching a decision, including information from the user, from Meta, outside experts and public commenters. The panel then releases a draft decision which is communicated to all OB members; a majority of its members must be reached for the decision to be approved and publicly released. Decisions can rule to uphold or overturn the Meta decision being discussed; it can also provide interpretation and recommendation on Meta's policies, standards and procedure. Then, Meta has to implement the OB's ruling and respond within sixty days to the policy recommendations included in the decision, even if it has no obligation to implement the recommendations. Similar to case law, the OB's decision can be taken into account for deciding on future decisions. When it comes to independence, the OB is financed by an independent trust set up by Meta.

Critical analysis

Self-regulation from platforms on matters such as content moderation is only the logical follow-up to the evolution of the regulation of the online sphere. There is a growing trend in law and

[330] 'Governance | Oversight Board' <https://www.oversightboard.com/governance/> accessed 18 January 2023.
[331] 'Oversight Board Cases' (*Meta Transparency Centre*) <https://transparency.fb.com/oversight/oversight-board-cases/> accessed 20 March 2023.

policymaking asking more from platforms to protect fundamental rights including by the adoption of more stringent and precise legislation and obligations as demonstrated in Section 3 of this deliverable.

Interestingly, some authors pointed out that Facebook OB differs from other self-regulation initiatives.[332] Others[333] which are overseen by regulatory institutions while the OB is independent of any regulatory oversight. Therefore, this type of self-regulation can be interesting to avoid States being too heavily involved in speech regulation and avoid potential censorship. It remains to be seen if it is truly impactful or if the enforcement aspect is too low.

The OB is a controversial institution and has both supporters and critics. In 2022, D. Wong and L. Floridi conducted a comprehensive overview of scholars' criticisms or praises about the OB that you can find below.[334] We highly recommend their work for a full overview and select below some relevant elements of their analysis.

Advantages

Transparency

Firstly, the OB has definitely improved the transparency of content moderation decisions and provided a recommendation to improve blur concepts in Meta's policies. It can highlight some loopholes and incite the revision of those such as the exemption to some content moderation rules for certain public figures following the Facebook files disclosure by the Wall Street Journal.[335]

Game changer

Many OB policy recommendations have been voluntarily implemented by Meta even if they are non-binding.[336] This format has been shown to provide flexibility to Meta to implement and deal with sensitive freedom of expression topics.[337]

[332] Cusumano, Gawer and Yoffie (n 329).

[333] Such as the Code of Conduct on countering illegal hate speech online or the Code of Practice on Disinformation

[334] David Wong and Luciano Floridi, 'Meta's Oversight Board: A Review and Critical Assessment' [2022] Minds and Machines <https://doi.org/10.1007/s11023-022-09613-x> accessed 21 December 2022.

[335] 'The Facebook Files' *Wall Street Journal* (1 October 2021) <https://www.wsj.com/articles/the-facebook-files-11631713039> accessed 18 January 2023; Elizabeth Dwoskin and Cat Zakrzewski, 'Facebook's Independent Oversight Board Demands Transparency on Exemptions for Politicians' *Washington Post* (21 September 2021) <https://www.washingtonpost.com/technology/2021/09/21/facebook-oversight-board-transparency-xcheck/> accessed 18 January 2023.

[336] Wong and Floridi (n 334).

[337] Evelyn Douek, 'Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility' (2019) 21 North Carolina Journal of Law & Technology 1.

Assertiveness

Statistics show that the OB is not afraid to overrule Meta decisions and exercise its powers.[338] The vast majority of cases actually overturn Meta's initial decisions. This doesn't guarantee the quality of decisions.

Shortcomings

Lack of diversity

The OB is composed of 23 members despite a male-female balance, a geographical balance is still lacking with a majority coming from the US. Especially as no co-chair comes from Africa or China, and they are the ones deciding on staff hiring matters, case selections, etc. Furthermore, the representation of LGBTQ+ and disabled communities on the board is also lacking. This imbalance could impair and impact the content moderation decisions and the future of the board. The geographical imbalance is also visible in the number of appeals submitted against Meta's decisions with a real underrepresentation of global south appeal applications.[339]

Limited impact

The OB has a limited mandate in the content it can review. Decisions on accounts or groups and features such as recommendation algorithms and advertising systems are excluded. However, Edward Pickup argues that the Charter and the Bylaws could actually be interpreted as enabling the OB to access Facebook's algorithms as part of its standard review process and to make recommendations regarding algorithms' impact on Facebook.[340] The OB cannot go beyond its jurisdiction.[341] The OB cannot review cases where a content moderation decision could lead to adverse governmental action against Meta such as content unlawful in the jurisdiction of the posting or reporting party. This limits the scope of freedom of expression questions emerging on Meta and Instagram through the OB. Douek argued that for the OB to be meaningfully empowered to review the main content moderation, it should not be assigned to review only a small subset of them that are peripheral to (Meta's) main product.[342] This selective approach also raises questions in terms of equality as not every user can benefit from the review. In addition, theoretically, Meta is not bound by the OB decisions and policy recommendations and they could choose to cut the OB funding.

[338] 'Oversight Board Publishes First Annual Report | Oversight Board' <https://www.oversightboard.com/news/322324590080612-oversight-board-publishes-first-annual-report/> accessed 19 January 2023.

[339] 'Oversight Board Publishes First Annual Report | Oversight Board' (n 338).

[340] Edward L Pickup, 'The Oversight Board's Dormant Power to Review Facebook's Algorithms' <https://openyls.law.yale.edu/handle/20.500.13051/18219> accessed 19 January 2023.

[341] Dipayan Ghosh, 'Facebook's Oversight Board Is Not Enough' [2019] *Harvard Business Review* <https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough> accessed 16 November 2022.

[342] Evelyn Douek, 'What Kind of Oversight Board Have You Given Us?' (2020) 2020 University of Chicago Law Review Online 1.

Future of the Oversight Board

As promptly noted by Wong and Floridi, the OB is presenting similarities with some newly adopted DSA obligations without matching their scope. For instance, now online platforms must set up an internal complaint-handling system (article 21 DSA). However, the OB's current shape doesn't match the obligation for several reasons. First of all, it is not an internal body, indeed the OB is legally independent of Meta. In addition, this internal complaint system would be available to anyone who is unhappy about the decision following content being notified without misuse (recurrent and abusive use of the protection granted in the DSA). The OB's scope of review is much more selective. Indeed, Aleksandra Kuczerawy underlines that during the Q2 of 2022 out of 347,304 cases submitted OB has only delivered decisions on three.[343] This means Meta will have to adapt the mandate and organisation of the OB or stick with it but develop internally such a complaint handling system.

In February 2023, the OB announced significant changes to their Charter and Bylaws to enable the OB to review more cases and to do so faster than before.[344] This important change will have a great impact over the cases reviewed. Indeed, so far in two years they took around 35 decisions.[345] They have always used standard decision but now the Board expressly mentioned that expedited decisions will be finally used providing a review from 48 hours to 30 days upon acceptation of the case by the Board.[346] In addition, more summary of cases decision will be published especially when Meta's decisions have been overturned. A review selection will be still operated which goes against the aim of the DSA internal complaint mechanism (art. 20) as not all complaints can go through this channel.

About the likelihood to see the OB becoming a DSA out of court dispute settlement, it appears that the OB would need to go through significant changes to be able to qualify as such.[347] The OB's mandate, functioning and scope are not matching the DSA scope. The OB would need to non-exclusively review Meta's moderation decisions, be financially distinct from Meta, broaden

[343] Aleksandra Kuczerawy, 'Social Media Councils under the DSA: a path to individual error correction at scale?', in: M. Kettemann (ed.), Platform://Democracy Project - Research Clinic Europe, commissioned by the Stiftung Mercator, and it is carried out by the Leibniz Institute for Media Research | Hans-Bredow-Institut (HBI) with support from the Humboldt Institute for Internet and Society (Berlin) and the Department of Theory and Future of Law of the University of Innsbruck (Austria). See more information https://leibniz-hbi.de/en/news/platform-councils-as-tools-to-democratize-hybrid-online-orders, 2023, forthcoming.

[344] 'Oversight Board Announces Plans to Review More Cases, and Appoints a New Board Member' <https://www.oversightboard.com/news/943702317007222-oversight-board-announces-plans-to-review-more-cases-and-appoints-a-new-board-member/> accessed 20 March 2023

[345] Oversight Board Announces Plans (n 344)

[346] Oversight Board Announces Plans (n 344)

[347] Aleksandra Kuczerawy, 'Social Media Councils under the DSA: a path to individual error correction at scale?', (n 343).

the scope of moderation decisions selected for review (not only focus on removal or non-removal of content), get way more human resources, ….[348]

As concluded by Wong and Floridi, the OB has shown its current limits but still shown the potential to be much more. If improved, the OB could become a "valuable complement to robust, international legislation".[349,313] Since the creation of the Facebook Oversight Board, other platforms have followed and embraced the path of self-regulation models for their content moderation processes and decisions. We can name the Twitter Trust and Safety Council, the TikTok Content Advisory Council, the Spotify Safety Advisory Council, and Twitch's Safety Advisory Council.

4.2.2 Social Media Councils

In 2018, Article 19,[350] suggested exploring a new model of effective self-regulation for social media. This model would contain social media councils (SMC).[351] While the SMC are relatively new, the underlying idea is not, as they are highly inspired by the press/journalist councils, long-established self-regulation bodies for the press and journalists.[352] They are for instance leading the way for journalistic deontological code, ethical code, and so forth. The idea behind their creation was to solve some content moderation issues on social media such as hate speech. The purpose is also to establish conditions for "independence, openness to civil society participation, accountability and effectiveness"[353]. These Social Media Councils would become a multi-stakeholder, transparent, inclusive accountability mechanism for content moderation on social media. Social Media Councils would make sure that "decisions on content moderation are compatible with the requirements of international human rights standards and are shaped by a diverse range of expertise and perspectives".[354]

Since then, Article 19 has consulted a wide range of actors and set up a pilot in Ireland.[355] The civil society organisation has also shaped better its proposal for Social Media Councils.

[348] ibid.
[349] Wong and Floridi (n 304).
[350] Article19 is a leading free speech global organization.
[351] 'Self-Regulation and "Hate Speech" on Social Media Platforms' (*ARTICLE 19*, 2 March 2018) <https://www.article19.org/resources/self-regulation-hate-speech-social-media-platforms/> accessed 21 February 2023.
[352] Stefanie Barth, 'Can Social Media Councils Tame Digital Platforms?– Digital Society Blog' (*HIIG*, 29 September 2022) <https://www.hiig.de/en/social-media-councils/> accessed 1 March 2023.
[353] 'Self-Regulation and "Hate Speech" on Social Media Platforms' (n 351).
[354] 'Social Media Councils' (*ARTICLE 19*) <https://www.article19.org/social-media-councils/> accessed 21 February 2023.
[355] Article 19, 'Social Media Councils, One Piece in the Puzzle of Content Moderation' (2021) <https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf>.

Article 19 sets up the key objectives of SMC, as follows:

“Key objectives

- **Review individual content moderation decisions** made by social media platforms on the basis of international standards on freedom of expression and other fundamental rights.
- **Provide general guidance** on content moderation guided by international standards on freedom of expression and other fundamental rights.
- Act as a **forum** where all stakeholders can discuss and adopt recommendations or interpretations.
- Use a **voluntary-compliance approach** to the oversight of content moderation.”

Source: Article 19, ‘Social Media Councils, One Piece in the Puzzle of Content Moderation’ (2021).

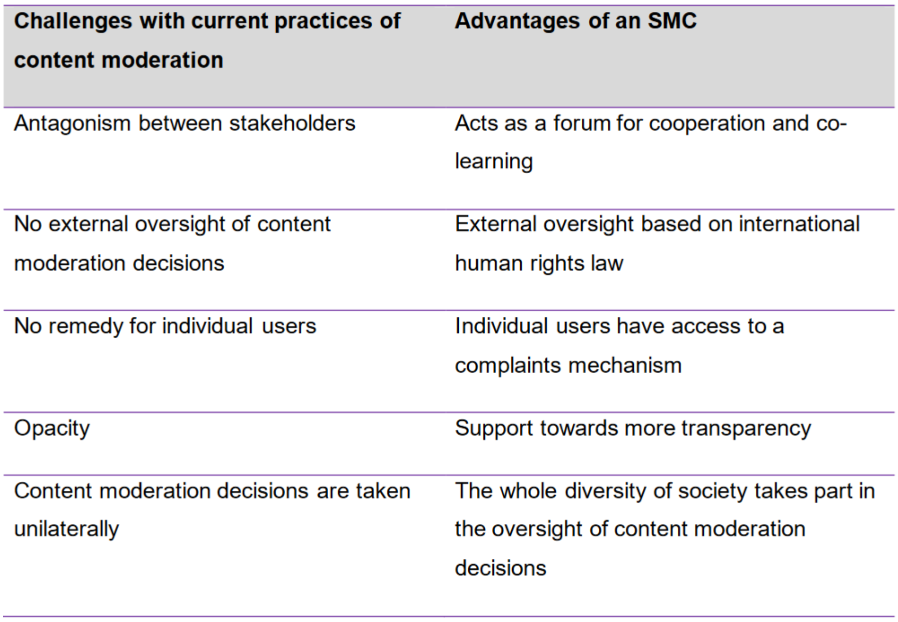
Based on the expertise and experience of Press Councils, the SMC should follow the following principles:

1. “Be independent of government, commercial, and special interests.
2. Be established via a fully consultative and inclusive process – major constitutive elements of their work should be discussed in an open, transparent, and participatory manner that allows for broad public consultation
3. Be democratic and transparent in their selection of members and decision-making.
4. Include broad representation – it is important that the self-regulatory body includes representatives that reflect the diversity of society (including the representation of minorities and groups in situations of vulnerability or marginalisation).
5. Have a robust complaint mechanism and clear procedural rules to determine if applicable standards were breached in individual cases.
6. Have the power to impose only non-financial sanctions.
7. Work in the public interest and be transparent and accountable to the public”.

Source: Article 19, ‘Social Media Councils, One Piece in the Puzzle of Content Moderation’ (2021).

The Social Media Council concept has actually been endorsed by the UN special rapporteur on the promotion and protection of the right to freedom of opinion and expression, namely David Kaye.[356]

Among academics, civil society, and private stakeholders some advantages and challenges of the SMC model were outlined and summarised. The overview can be found below in a table from Article 19, 'Social Media Councils, One Piece in the Puzzle of Content Moderation' (2021) (Figure 5).



Challenges with current practices of content moderation	Advantages of an SMC
Antagonism between stakeholders	Acts as a forum for cooperation and co-learning
No external oversight of content moderation decisions	External oversight based on international human rights law
No remedy for individual users	Individual users have access to a complaints mechanism
Opacity	Support towards more transparency
Content moderation decisions are taken unilaterally	The whole diversity of society takes part in the oversight of content moderation decisions

Figure 5: An overview of content moderation challenges and the advantages of the Social Media Councils model
Figure source: Article 19, 'Social Media Councils, One Piece in the Puzzle of Content Moderation' (2021)

Rules coming from the SMC infrastructure would come from the international and human rights framework. Social media companies "should ensure that their terms of service comply with international standards on freedom of expression as a consequence of their responsibility under the UN's Guiding Principles on Business and Human Rights."[357] Companies would have to respect and execute the SMC's decisions (or recommendations) in good faith. In relation to the choice of rules determined to govern the decisions of the SMC, some have raised the idea to create a code of human rights principles for content moderation. Other points of discussion are about

[356] University of Stanford, Global Digital Policy Incubator, Cyber Policy Center, 'Social Media Councils: From Concept to Reality - Conference Report' <https://cyber.fsi.stanford.edu/gdpi/content/social-media-councils-concept-reality-conference-report> accessed 21 February 2023.

[357] Article 19 (n 355) 19.

the role of the SMC, strictly advisory or adjudicatory, and the interplay between a global SMC and locally rooted SMC with their cultural and linguistic specificities.[358]

SMC and DSA

A. Kuczerawy gives a closer look to the SMC and the DSA new regime.[359] She investigates whether these models could be considered as an internal complaint mechanism or an out-of-court dispute settlement. SMC will not constitute an internal complaint mechanism given their external character. Indeed, as the name indicates, these mechanisms must be internal, meaning associated, financed and following the policies and rules decided by the platform. The purpose of SMC is to have multiple stakeholders on board. However, SMC could be a better fit as an out-off-court dispute settlement. The important aspect is that members of the out-of-court dispute settlement must be independent and impartial (art. 21).[360] Indeed, the condition of independence is a key aspect of the DSA provision on alternative dispute settlement as it provides legitimacy to the decisions delivered even if non-binding.[361] For more information on this topic, we refer to the analysis of A. Kuczerawy.[362]

In conclusion, “neither pure self-regulation nor aggressive government regulation seems likely to cover all the challenges digital platforms face”.[363] While the Facebook Oversight Board remains a step forward in the landscape of social media content moderation, it is only a company’s initiative for its own services. Social Media Councils bring the promises of multi-stakeholder voluntary compliance. They “would not solve the structural problems in the platforms’ business models, but they could offer an interim solution, an immediate way to start addressing the pressing problems in content moderation” and bring the societal aspects of content moderation to the forefront.[364] What the authors observed is that for self-regulation to be effective, it cannot happen at the exclusive firm level. Indeed, coalitions of firms within the same market and with similar business models may agree to abide by a jointly accepted set of rules or codes of conduct.[365] Most firms will not self-regulate without government pressure;[366] we can add without public scrutiny pressure as well. The issue is that governments and public authorities often do not have the resources (financial, human or expertise) “to regulate and monitor the dynamic, ongoing changes inevitable with digital platforms and their complex

[358] Pierre-François Docquir, ‘The Social Media Council: Bringing Human Rights Standards to Content Moderation on Social Media’, *Models for Platform Governance* (2019).

[359] Aleksandra Kuczerawy, ‘Social Media Councils under the DSA: a path to individual error correction at scale?’, (n 343).

[360] ibid.

[361] ibid.

[362] ibid.

[363] Cusumano, Gawer and Yoffie (n 329).

[364] Heidi Tworek, ‘Social Media Councils’, *Models for Platform Governance* (2019).

[365] Cusumano, Gawer and Yoffie (n 329).

[366] Jr Kwoka and Tommaso M Valletti, ‘Scrambled Eggs and Paralyzed Policy: Breaking Up Consummated Mergers and Dominant Firms’ <https://papers.ssrn.com/abstract=3736613> accessed 1 March 2023.

technologies and operations"[367]. Research showed that to solve these challenges and ensure efficient self-regulation, private and public actors will have to collaborate. We have seen this trend in the negotiation of the revised Code of practice on disinformation shifting from a self-regulatory instrument to a co-regulatory instrument and in the negotiation of the new DSA legislation. Further research questions on the framework for collaboration and incentives for successful regulatory initiatives remain to be answered.

[367] Cusumano, Gawer and Yoffie (n 329).

5 AI4Media Workshop on AI and content moderation

On Monday, 6 February 2023, KUL and UvA organised a workshop on AI and content moderation. AI4Media aims to explore concrete challenges faced by the industry, such as how to evaluate recommender systems, how AI can be used in audio-visual archives and, finally, the use of AI in content moderation. It is in relation to this last challenge that the team decided to organise a workshop with practitioners to discuss what are the main challenges faced by those either building AI systems for content moderation or using these systems. The purpose was to open up the discussion and learn more from their respective experience on the use of AI systems assisting their content moderation efforts. The workshop was limited to the Western perspective on content moderation, mostly EU.

The workshop was held online under the Chatham House Rule and was an invitation-only event to have a high engagement and a fruitful discussion. The workshop was an opportunity to network and discuss critical questions with the industry. Beyond this, participants were informed that the outcomes of the workshop would inform our ongoing research in the project as well as feed into this content moderation report.

Participants
A diverse and engaging team of participants expressed their interest for the workshop. The following actors participated in the workshop:

- A European company producing image recognition solutions for developers and businesses.
- A European consultancy doing content moderation analysis.
- An AI4Media-funded project focusing on robust and adaptable comment filtering.
- A prominent newspaper from Austria.
- A UK company developing socially Responsible AI for Online Safety. They develop AI-powered tools to find and stop toxic content.
- A German local broadcast media production and distribution company doing responsible journalism and professional entertainment.
- An American technology company that owns a very large online platform(s).
- A European company developing trustworthy, transparent and explainable human-centred AI solutions that read and understand large amounts of texts.
- A researcher from a well-known university in the Netherlands and consultant for the United Nations' Department of Political and Peacebuilding Affairs (DPPA) Innovation Cell.

Prior to the meeting, participants were asked to fill in a short survey asking to identify the top three challenges they are currently battling with in their daily work on AI in content moderation. This enabled structuring the discussion in advance and preparing relevant questions.

Speaker and critical discussants

Beyond the participants, the workshop welcomed the Distinguished University Professor of Law & Digital Technology, with a special focus on AI, Natali Helberger from the University of Amsterdam (UvA). She gave a short introductory talk on the regulatory landscape. Two discussants were also present to help and guide the discussion in the second part of the workshop: Bernhard Rieder, Associate professor in New Media and Digital Culture at the UvA and Aleksandra Kuczerawy, postdoctoral researcher at KU Leuven focusing on online Content Moderation and the Rule of Law.

Agenda

On the day of the workshop, first an introductory talk on the legal aspects of content moderation was given by Prof. Natali Helberger. Afterwards, a roundtable was held where each invited speaker briefly (5 minutes) shared what they considered their main challenge when working with AI-enabled content moderation (technical, economic, ethical challenge, etc.). The organisers then quickly summarised the **main challenges** outlined by the participants. Based on these findings, the group dived into the best practices discussion. In this discussion, the aim was both to narrow down more precisely what is producing these challenges and also discuss best practices in how to solve or mitigate them. Here how policy could potentially support these efforts was also explored. Bernhard Rieder and Aleksandra Kuczerawy provided their critical perspectives to support a fruitful discussion. Finally, the next steps were discussed with the participants. The participants were eager to stay in touch about further events and contribution opportunities in relation to AI4Media.

In addition to the workshop, a private session was held with a team from a European company that provides content moderation solutions for a wide range of clients, as they couldn't attend the workshop. They were extremely eager to discuss these topics and provide information about their experience.

Challenges

The main challenges identified by the workshop participants were the following:

1. Lack of access to training data to have accurate output (e.g., from social media platforms, open data sets, multiple languages).
2. Lack of transparency in:
   a) how the AI models are trained and maintained
   b) how the AI models are evaluated

c) how the use of AI models is disclosed to users.

3. Ensuring human oversight and real-time moderation.
4. Defining and classifying ‘hate speech’, ‘toxicity’ in a context-sensitive way, language contexts, intention, etc.
5. Inclusivity: minor languages are not properly represented.

Takeaways

The main takeaways from the workshop and our bilateral discussion are summarised below:

- AI is only a tool at human’s disposal. There is a lot of misunderstanding about what AI is and what AI can do on content moderation. AI content moderation systems should keep the human review component, since it’s extremely dangerous to fully automate the content moderation task. AI can be there to ease the task but it must not replace a human interpretation.
- The impact of content moderation on the human reviewers should not be under-estimated. It should be taken into account in a more effective way in future initiatives.
- Often, the same models are used for the architecture and design of AI models used for content moderation. Innovation on this aspect is complex. More attention and efforts could be allocated on fine-tuning the models and learning how to reduce the noise.
- Evaluation methods, processes and criteria for the models should be established in order to evaluate the positive and negative impacts of the models.
- The dislocation of content moderation is a worrying aspect. For instance, when content is being removed in some countries but not in others based on contextual and linguistic interpretation. In addition, there is a risk for content to be either under-moderated (not being detected and moderated) or over-moderated (massively moderated) based on the values established in the legal and policy document on the platforms. Some have raised the attention around concerns of data colonialism, the normative values of the judgement and the dominant views on content moderation. This could happen when only one community and geographical region of the world is setting the terms for the platform operating in all parts of the world.
- The work of human rights workers, archivists, historians should not be forgotten. Some abusive content can be valuable for documentation purposes useful for historical or research purposes, legal action, memory duty, ...
- The online space has taken such a considerable place in society that it leads to a question about access to information by the public. Several participants have raised the following question: could the content which is removed be considered as public domain information? The content removed could be part of the democratic debate. Could a right to request already moderated data be established?
- The respect to the GDPR is often used as an excuse not to share the data on removed content.

- The debate about content moderation should be open to a wide range of actors: small, mid and big players and at the different levels of the chain.
- The size of the company is an important aspect to look at, as for small and midsize companies the workload gets very hard to handle and then big problems might slip too. It also could really hurt the start-ups and prevent them from operating.
- An advisory board on content moderation and AI would be welcome to be a contact point to better listen to the variety of players active in content moderation.
- It was brought to our attention that some content moderation subjects are overlooked such as fraud, direct incitement to violence, self-harm, and crime plotting.

The rest of the results of the workshop have been already integrated in specific sections of this report.

6 Policy recommendation on content moderation

The research conducted for this deliverable showed how content moderation challenges can be specific but interconnected at the same time. This peculiar situation at the crossroad of freedom of expression and other fundamental rights makes it a complex topic to regulate. Content moderation follows a constant balancing exercise in a multi-layered and complex infrastructure and institutional landscape. This section will investigate the takeaways of this research and inform several policy recommendations for content moderation. It will first sketch some general and horizontal observations before diving specifically in each sub-content section as addressed in the mapping of the EU content moderation landscape.

6.1 Horizontal and high-level recommendations

Content moderation is not new. It varies from editorial screening, curation and organisation of content in libraries, schools, and media, to censorship and freedom of expression or copyright infringements cases. Content moderation has been part of the regulatory landscape for some time. However, the scale of it - thanks to the online space, the growing power of big tech actors and user-generated content has lifted the content moderation discussion to important instances and pushed to rethink our approach to free speech. Jennifer Holt talks about regulatory hangover to refer to the "conceptual and practical inability of policy to keep pace and evolve with technological and cultural change, resulting in media and communications infrastructure(s) veering off their foundational regulatory paradigms".[368] It is true that the scale, dynamic and challenges of online content moderation invite us to rethink power allocation and accountability mechanisms. Indeed, the shift of power from the traditional public authority to private actors to manage the public debate is reshaping the traditional approach to content moderation.

Content moderation trade-offs

Content moderation challenges are a moving target due to contextual and linguistic nuances. For instance, a certain word can have different meanings depending on the culture or a particular community or language using it, and it could also have a totally different meaning following certain events. In this deliverable, we observed how content moderation regulation is an outcome of complex, social-political decisions. The perfect solution to solve all the challenges expressed does not exist. It will always be a matter of trade-off and delicate balance of the various interests and fundamental rights at stake. The question of normative values in content moderation decisions is very much context and topic-specific.

Content moderation is located at the crossroad of several fundamental rights, including the freedom to conduct business, freedom of expression and the EU Member States' positive

[368] Jennifer Holt, 'Cloud Policy: Anatomy of a Regulatory Crisis' (October 2017).

obligations under the fundamental rights law framework. Jialun 'Aaron' Jiang identified several dilemmas/trade-offs when it comes to content moderation[369]:

- **Transparency vs. Security**: While providing transparency can greatly improve accountability legitimacy, certainty about the behaviour to adapt it can also be misused by malicious actors to circumvent rules, detection techniques and so forth.
- **Rapidity vs. Accuracy**: On some occasions, content needs to be immediately taken down; the faster the removal is being operated, the more the context and nuances cannot always be fully appreciated, impacting the quality and accuracy of the removal.
- **Nurturing vs. Punishing**: Different content moderation policy approaches can be used, whether that would be an educational approach to improve and reform certain behavior online (nurturing), or the punitive approach that bets on sanctions for rules violations. Both approaches have their advantages and disadvantages and a combination of the two has a stronger effect than each in isolation.
- **Quantity vs. Quality**: Intermediary services providers can be caught between two conflicting interests: having a lot of content, traffic and engagement or having some high-level quality content hosted on their services. Different values and interests are at stake here.

Which regulatory approach for content moderation?

This deliverable concludes that neither the strict and binding framework of the law nor the self-regulation initiatives or end-user moderation are providing alone the answer to content moderation challenges. Indeed, voluntary initiatives, such as Codes of Conduct or Codes of Practice, have been deemed unsatisfactory or raising the question of accountability and legitimacy. They are often a good complement to legislation or a starting point to move content moderation initiatives forward whether it would be sectoral soft law or hard law.

On the other hand, the strict legislative approach can appear to not be easily adapted for solving a moving target or, on the contrary, well-designed but unenforced. The analysis in Section 3.2 of this deliverable allows to identify some gaps and shortcomings regarding the European framework for the liability and responsibilities of hosting service providers. For more specific recommendations, please consult the next section, 6.2.

Firstly, there is an increasingly fragmented content moderation landscape. This shows indeed that the approach chosen is to follow a content specific approach. In the last few years, existing legislation has been adapted or new ones have been adopted, making this area of law a hot topic. National[370] and European regulators have decided to react to the content moderation challenges with a more proactive approach in order to ensure a safe online environment, which is respectful of competition rules and EU values. This regulatory effort has created a regulatory

[369] Jialun 'Aaron' Jiang (n 32).

[370] Such as Germany with the adoption of the NetzDG law and the United Kingdom (ex-EU MS) with the Online Safety Bill.

"jungle" for the general public or for the ones that cannot afford legal advice or afford the time to investigate full compliance. There is a strong need to take the landscape as a whole and make sure all the specific legislations are addressing the specific needs of their targets while still working together in the bigger picture. In a single message, one can have terrorist content and homophobia (hate speech) triggering different rules and legal mechanisms.

This deliverable identified three main categories of inconsistencies:

1. The lack of consistent implementation of the legal framework;
2. The lack of a comprehensive and harmonised definition of what counts as illegal content;
3. Diverging measures between the various instruments of the legal framework.

This deliverable also identified five gaps and missing safeguards in the regulatory landscape.

1. The lack of clarity concerning the criterion of actual knowledge;
2. The lack of clarity concerning the scope of procedural obligations to limit the dissemination and the applications of sanctions;
3. The lack of adequate safeguards for fundamental rights;
4. The lack of agreement as to what constitutes adequate procedural measures;
5. High compliance costs for SMEs.

Some of these gaps have now been addressed by the Digital Services Act. However, when it comes to content moderation legislation, a considerable challenge is about the enforcement of the legal provisions. For instance, as analysed in this document, many provisions exist already, before the DSA was adopted, and these rights and obligations should be known and respected. The coordination and collaboration between the various authorities responsible for the different layers of enforcement is an extremely important aspect of a successful enforcement strategy. For instance, for the DSA, there is an entanglement of different actors involved in the enforcement: the Digital Service Coordinators established in each EU Member State, the European Board of Digital Services, the European Commission, and finally, the national courts. How this will work in practice remains to be seen. It must also be ensured that the challenges which augmented and virtual reality will create for content moderation, fall within the scope of the current laws.

A wholesome policy approach is needed

However, not every challenge on content moderation can be solved by setting up self or hard regulation. The crucial role of education and literacy initiatives cannot be underestimated. From a young age, education about user-generated content rights and duties, the type of illegal and harmful content, freedom of expression limits and the various options of content moderation and remedies available are important cornerstones of a more safe and trustworthy online environment.

The choice of the technology and content moderation approach used also matters a lot. Not all technology will produce the same results of efficiency for all types of content. It is therefore important to have a careful look at the approach supported and chosen. For instance, content moderation of voice, image, text and live stream content requires different visions. Indeed, in live streams, it's not always easy to act swiftly especially as the moderator needs to have evidence someone broke a rule or find a technology to delete or annotate someone's voice as they speak.[371]

AI content moderation systems: powerful tools needing careful framework and caution

When it comes to AI systems, it is important to underline how AI is an efficient and often needed tool, but it is not a perfect tool. It is important to remain realistic about what AI can achieve and be aware of the great promises and risks. The adverse effects when it comes to content moderation are not hypothetical, and in the short term, they are already impacting online content, and can in the mid-long term be even more severe. Whether it would be self-censorship, privacy infringements or restrictions on legal content, great care is required when AI systems enter the media sector in light of its crucial role for democracy as a place for individual expressions. It is important to remember that AI systems for content moderation purposes are designed, shaped, managed and supervised by human beings. It is therefore tremendously important to bring ethical and legal considerations on the table of AI content moderation systems development.

There is a need to build bridges across different sectors or internal departments to make sure to have a fair, balanced and sound approach to AI dealing with content. Nuances, languages and context are key when it comes to content moderation and AI systems are far from being able to completely take over the human interpretation of content, but they can greatly help the review work in light of the scale.

Inclusivity and diversity

Geographical and diverse communities should be better considered when talking about moderating content. There seems to be a lack of diversity not only on the gender basis but also on the geographical and ethnic representation for developing AI systems, terms and conditions and taking content moderation decisions by a human. As all is a matter of contextual interpretation with content moderation, people from different regions will perceive abusive behaviour differently. Some pointed out that using a single set of rules to regulate global users falsely implies that people view abusive behaviour consistently across the world.[372] Having community guidelines from a predominantly Western, and more specifically U.S. perspective, raises issues of inclusivity and relevance for different content contexts and points of view. "The

[371] Jialun 'Aaron' Jiang (n 32).
[372] Jialun 'Aaron' Jiang (n 32).

consistencies and variances across the world in the perceptions of abuse reveal the complexity of content moderation, and the necessity of a multi-stakeholder perspective".[373] For instance, local experts should be consulted and critically consider the local meanings of community guidelines when they are translated. [374]

Multi Stakeholder consultation

While combining and adapting technologies, approaches and regulations to get the best content moderation approach for certain content and context, there is the same need to ensure a multi-stakeholder consultation and involvement to improve the quality of the content moderation. Better empowerment of end-users and civil society is necessary to improve content moderation efforts and shift the power asymmetry. Users whose rights are affected need to step in. The DSA brings new obligations and rights which will improve this aspect. There are various new empowering instruments established, such as notice and action procedures, internal complaint-handling mechanisms and trusted flaggers. The enforcement and literacy around these aspects will be crucial. Users should be made aware of these new rights and make the most of them. Given the expertise of civil society organizations, they should be encouraged, perhaps more institutionally, to work together with the EC to better oversee VLOPs.

Human role in content moderation

There is a need for greater role and transparency about human involvement in the content moderation process. Whether it would be transparency about the working conditions, training and processes for content moderators, or the impact of their work on their health, there is a need to have more protection in place. Human moderators have a crucial role in shaping the online debate, great care should, therefore, be brought about their position. The moderators' skills should match the content they moderate. This includes knowledge of language of the content and socio-political nuances of the context of a given content. Relevant training and sensitivity vis-à-vis fundamental human rights at stake are required when moderating online speech. The recruitment of sub-contracting parties for content moderation purposes is not always clearly mentioned, and a lot still unknown about this emerging work market.

Research needs

Content moderation is a subject for which more information is needed from private actors. Black boxing corresponds to situations where a private company has aggressive control over information on technical infrastructure, business operations or labour practices.[375] Indeed, not so much is known about the infrastructure but also whether some archive centre for removed content exists and if they can be accessed. The research community and civil society are, as a matter of fact, concerned about not having access to information about the content being

[373] Jialun 'Aaron' Jiang (n 32).

[374] Jialun 'Aaron' Jiang (n 32).

[375] Parks (n 13).

removed. This includes content that could be used for historic, research, and archive purposes or for collecting evidence for a legal case. The question remains whether these data are then being used to train the algorithm or if there is a human team reviewing the process. Hopefully via the DSA and Digital Markets Act[376] more information on these aspects will be provided thanks to the transparency reports, audits, access requests and so forth. However, there is the risk that a lot will be kept secret for privacy, IP, trade secrets or security grounds.

Conclusions

Content moderation will most likely be always governed by unresolvable tensions between competing interests and conflicting fundamental rights. There will not be a magic formula to clear all hosting platforms from illegal or harmful content. A combination of technologies, regulatory approaches, contextual interpretation and multi-stakeholders' consultation is needed to achieve a balanced approach.

In the following, we summarise the high-level policy recommendations for content moderation, based on the analysis presented in the previous paragraphs of Section 6.1.

- Envisage a combination of regulatory instruments, technologies and content moderation approaches to fit the specificities of context and content.
- Ensure proper communication, awareness raising and compliance support about the complex EU regulatory landscape (targeting end-users, small and mid-field players).
- Ensure consistency between the various content moderation legal instruments on their intersection aspects.
- Investigate which technologies and approaches work best for what type of content and context.
- Take into account geographical location, languages and diverse communities for various aspects of content moderation.
- Tailor the use of the technology and the approach chosen in light of the content being moderated (text, image, live stream, etc.).
- Ensure the regular updates of the terms of use, community guidelines in light of the constant evolution of content moderation.
- Ensure proper training, expertise, and skills for human moderators in light of the content they moderate.
- Ensure more transparency and safeguards about content moderation sub-contracting and working conditions of human moderators.
- Improve the transparency about the content moderation infrastructure and data (deletion, archive, transfer).

[376] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) 2022 (OJ L).

- Ensure proper processes in place for research, historical, archival, lawsuit purposes by specific actors.
- Ensure the enforcement of the existing and new tech legislations impacting content moderation such as the empowerment, transparency and access provisions in the DSA and DMA. This will improve content moderation efforts and avoid black boxing, ensure accountability and enable a better understanding of content moderation mechanisms and unidentified challenges.
- Ensure a proper balance between AI systems and human moderation.
- Empower content moderation stakeholders: end-users, civil society, researchers, historians, archivists, etc.

6.2 Problem-specific recommendations

In addition to the horizontal and high-level recommendations of Section 6.1, this section presents a set of problem-specific recommendations focusing on specific types of content such as terrorist content, copyright-protected content, child sexual abuse material, hate speech, and disinformation.

The problem-specific recommendations can be found below in Table 1.

Table 1: Problem-specific recommendations

NAME	RECOMMENDATIONS
Terrorist content	○ Re-consider the 1-hour window for action upon order receipt. ○ Discourage platforms from using voluntary ex ante upload filters. ○ Consider a different set of obligations for hosting providers of smaller size or reach of the service. ○ Ensure independent judicial review for takedown orders. ○ Enhance greater transparency of public and private collaboration.
Copyright-protected content	○ Discourage platforms from using voluntary ex ante upload filters, while ensuring human review for ex ante removals. ○ Allow ex-ante upload filters only for manifestly infringing content. ○ Strengthen ex post human review for removed or blocked content. ○ Establish more efficient reinstatement and redress mechanisms for erroneous removals. ○ Ensure effective and simple counter-notice processes. ○ Encourage platforms to adopt preventive policies safeguarding removal of work in the public domain or work benefitting from a non-exclusive license, exceptions, or limitations.

NAME	RECOMMENDATIONS
	○ Consider creating a centralised repository of public domain and non-exclusive licensed works where platforms could benefit from for their ex-ante reviews, as well as allow legitimate uses to avoid unreasonable removals or blockings.
Child sexual abuse material	○ Develop literacy initiatives to empower and educate children and teenagers about CSAM and their rights in light of the new legislation. ○ Conduct wider tests on the technologies available to achieve the moderation policy goals. ○ Consider all the possible channels for CSAM to circulate on intermediary services providers in order to adapt sound and relevant strategies and adequate legal provisions. ○ Ensure transparency of collaboration and processes for data exchanges between the relevant departments in charge of CSAM fight. Elaborate safeguards to frame cautiously the scope, and methods of the collaboration. ○ Conduct a careful balance assessment of the trade-offs between privacy/data protection and the objective to stop CSAM content.
Hate speech	○ Enhance transparency of the reporting systems to include information explaining, for example, which percentage of the removed content was found illegal after review. ○ Re-consider the 24-hours window for take down of "illegal hate speech".
Disinformation	○ Provide clear terminologies and definitions regarding the concepts mentioned in the Code of Practice on Disinformation. ○ Encourage non-VLOPs to become signatories of the Code and clarify their compliance and commitments. ○ Clarify the relationship between the DSA and the Code. ○ Assign an independent body with more resources and expertise to monitor compliance of signatories with the Code.

7 Conclusions

This deliverable described the efforts from policymakers to catch up with the growing power of intermediary services providers including big tech platforms over the online space, the scale of content and the use of technology to manage it all. The EU and national policymakers have adopted regulations on various aspects of content moderation. This plethora of legislative and non-legislative initiatives, policies, and laws raises challenges for synergies and coherence between various texts. Importantly, the delicate balance between conflicting interests or clashing fundamental rights is one of the aims which content moderation regulation is trying to achieve.

This deliverable also presented diverse approaches by different private actors (e.g. social media platforms) to address in their own way the challenges of content moderation on their services. Whether that would be by relying on end-users to moderate content or establish an independent board to oversee a selection of cases or exploring new self-regulation models for this growing tech market. Each of these approaches has its advantages and disadvantages.

In light of the aspects explored in this deliverable, it can be concluded that there is not a single way to address the multi-complexity of content moderation. Most probably, content moderation efforts are going towards a bundle of components for content moderation purposes. An encompassing approach would guarantee to make sure the specificities of the various types of content, actors and services are taken into account in content moderation decisions. The one size fits all approach does not match the issues encountered with content moderation even if a foundation of shared principles and safeguards is necessary.

Perhaps the future of content moderation will involve a more active role for end-users in the features they use in online spaces. For instance, the Digital Markets Act should ensure interoperability of online services on gatekeeper platforms; this will probably open the door to new opportunities for content moderation services. Platforms could see the emergence of new plug-ins or in-house features chosen by end-users to ensure accountability of content moderation decisions. The Digital Services Act has also granted users new procedural rights and imposed on online intermediaries a range of obligations. It remains to be seen how they are complied with, enforced by national authorities, interpreted by national and European courts.

With new technological advances, come new benefits, but also potential new risks for fundamental rights. As showed in this deliverable, this is the case for virtual spaces such as metaverse. How to reconcile an efficient removal of new forms of illegal and or unwanted content with fundamental rights of end-users (such as a right to privacy, freedom of expression) is becoming a pressing issue for content moderation regulation.

Shadow zone still exists in the content moderation sector, preventing sound analysis of challenges and potential remedies. This is the case either because of the platforms' secrecy, or because of the lack of access to data. It is, therefore, important to broaden the transparency on those aspects (institutional, infrastructure, work market, less represented type of illegal/harmful content). More research will be necessary to ensure that the fast-evolving content moderation initiatives (legal or not) are designed to balance all the values, rights and interests at stake. The adverse effects of content moderation on the mid and long-term for media, society and democracy are not yet known and should be carefully considered to ensure a sustainable online future.

8 References

Academic works, blogs, news, websites

- ‘Accessing Information about Abortion – Verfassungsblog’ <https://verfassungsblog.de/accessing-information-about-abortion/> accessed 7 March 2023
- Al-Waheidi RB Majd, ‘A Factory Line of Terrors: TikTok’s African Content Moderators Complain They Were Treated like Robots, Reviewing Videos of Suicide and Animal Cruelty for Less than $3 an Hour.’ (*Business Insider*) <https://www.businessinsider.com/tiktoks-african-factory-line-of-terrors-2022-7> accessed 23 February 2023
- ‘“Algospeak” Is Changing Our Language in Real Time - The Washington Post’ <https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/> accessed 21 February 2023
- Amnesty International, ‘Myanmar: Facebook’s Systems Promoted Violence against Rohingya; Meta Owes Reparations – New Report’ (Amnesty International, 29 September 2022) <https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/> accessed 17 March 2023
- Angelopoulos C and Senftleben M, ‘An Endless Odyssey? Content Moderation Without General Content Monitoring Obligations’ <https://papers.ssrn.com/abstract=3871916> accessed 1 February 2023
- Arsht A and Etcovitch, ‘The Human Cost of Online Content Moderation’ (*Harvard Journal of Law & Technology*, 2 March 2018) <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation> accessed 17 February 2023
- Article 19,
    - EU: Digital Services Act crisis response mechanism must honour human rights <https://www.article19.org/resources/eu-digital-services-act-crisis-response-must-respect-human-rights/> accessed 7 March 2023
    - ‘Joint Letter on European Commission Regulation on Online Terrorist Content’ (6 December 2018) <https://www.article19.org/resources/joint-letter-on-european-commission-regulation-on-online-terrorist-content/> accessed 9 February 2023
    - ‘Social Media Councils, One Piece in the Puzzle of Content Moderation’ (2021) < https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf>
- Barth S, ‘Can Social Media Councils Tame Digital Platforms?– Digital Society Blog’ (*HIIG*, 29 September 2022) <https://www.hiig.de/en/social-media-councils/> accessed 1 March 2023
- Bellanova R and de Goede M, ‘Co-Producing Security: Platform Content Moderation and European Security Integration’ (2022) 60 JCMS: Journal of Common Market Studies 1316

- Bertuzzi L,
  - ‘Ireland’s Privacy Watchdog Accused of Paralysing GDPR Enforcement’ (www.euractiv.com, 13 September 2021) <https://www.euractiv.com/section/data-protection/news/irelands-privacy-watchdog-accused-of-paralysing-gdpr-enforcement/> accessed 17 March 2023.
  - ‘Whistleblowers Are Impossible without Encryption, Edward Snowden Says’ (*www.euractiv.com*, 21 October 2021) <https://www.euractiv.com/section/data-protection/news/whistleblowers-are-impossible-without-encryption-edward-snowden-says/> accessed 25 January 2023
- Bishop B, ‘The Cleaners Is a Riveting Documentary about How Social Media Might Be Ruining the World’ (*The Verge*, 21 January 2018) <https://www.theverge.com/2018/1/21/16916380/sundance-2018-the-cleaners-movie-review-facebook-google-twitter> accessed 9 February 2023
- Bridy A. and Keller D., ‘U.S. Copyright Office Section 512 Study: Comments in Response to Notice of Inquiry’ [2016] SSRN Electronic Journal.
- Bukovská B, ‘The European Commission’s Code of Conduct for Countering Illegal Hate Speech Online’
- Buolamwini J and Gebru T, ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’ 15
- Business RI CNN, ‘Facebook Has Language Blind Spots around the World That Allow Hate Speech to Flourish’ (*CNN*) <https://www.cnn.com/2021/10/26/tech/facebook-papers-language-hate-speech-international/index.html> accessed 3 December 2021
- Butcher P, ‘Disinformation and Democracy: The Home Front in the Information War’ (*European Policy Centre*) <https://www.epc.eu/en/publications/Disinformation-and-democracy-The-home-front-in-the-information-war~21c294> accessed 31 January 2023
- Cambridge Consultants, ‘Use of AI in Online Content Moderation’ (2019) <https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation> accessed 20 February 2023
- ‘Caught in the Net: The Impact of Extremist Speech Regulations on Human Rights Content’ <https://syrianarchive.org/en/lost-found/impact-extremist-human-rights> accessed 1 December 2021
- Chandrasekharan E and others, ‘The Internet’s Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales’ (2018) 2 Proceedings of the ACM on Human-Computer Interaction 32:1
- ‘Commission Recommendation (EU) 2018/334 of 1 March 2018 on Measures to Effectively Tackle Illegal Content Online’, vol 063 (2018) <http://data.europa.eu/eli/reco/2018/334/oj/eng> accessed 1 February 2023

- ‘Community Standards Enforcement | Transparency Center’ <https://transparency.fb.com/data/community-standards-enforcement/> accessed 1 December 2021
- ‘Controversial Proposal on Combating Child Sexual Abuse Online’ <https://eucrim.eu/news/proposal-on-combating-child-sexual-abuse-online/> accessed 23 January 2023
- Cronin O, ‘An Garda Síochána Unlawfully Retains Files on Innocent People Who It Has Already Cleared of Producing or Sharing of Child Sex Abuse Material’ (*Irish Council for Civil Liberties*, 19 October 2022) <https://www.iccl.ie/news/an-garda-siochana-unlawfully-retains-files-on-innocent-people-who-it-has-already-cleared-of-producing-or-sharing-of-child-sex-abuse-material/> accessed 31 January 2023
- Cusumano MA, Gawer A and Yoffie DB, ‘Can Self-Regulation Save Digital Platforms?’ (2021) 30 Industrial and Corporate Change 1259
- De Blasio E. and Selva D., ‘Who Is Responsible for Disinformation? European Approaches to Social Platforms’ Accountability in the Post-Truth Era’ (2021) 65 American Behavioral Scientist 825.
- De Laat PB, ‘The Use of Software Tools and Autonomous Bots against Vandalism: Eroding Wikipedia’s Moral Order?’ (2015) 17 Ethics and Information Technology 175.
- De Streel A and Husovec M, ‘The E-Commerce Directive as the Cornerstone of the Internal Market’ [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3637961> accessed 24 January 2023
- De Streel A and others, *Study on Potential Policy Measures to Promote the Uptake and Use of AI in Belgium in Specific Economic Domains* (FPS Economy 2022)
- Dimitrov D., ‘DSA: Political Deal Done’ (Free Knowledge Advocacy Group EUApril 26, 2022) <https://wikimedia.brussels/dsa-political-deal-done/> accessed February 25, 2023.
- Directorate-General for Internal Policies of the Union (European Parliament) and others, *Online Platforms’ Moderation of Illegal Content Online: Laws, Practices and Options for Reform* (Publications Office of the European Union 2020) <https://data.europa.eu/doi/10.2861/831734> accessed 23 January 2023
- Discord Trust & Safety Team, ‘Digital Services Act - Information on Average Monthly Active Recipients in the European Union,’ <https://support.discord.com/hc/en-us/articles/12477677109143-Digital-Services-Act-Information-on-Average-Monthly-Active-Recipients-in-the-European-Union> accessed February 28, 2023.
- ‘Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis’ (*Center for Democracy and Technology*, 20 May 2021) <https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/> accessed 31 January 2023
- Docquir P-F, ‘The Social Media Council: Bringing Human Rights Standards to Content Moderation on Social Media’, *Models for Platform Governance* (2019)
- Douek E,

  - ‘Facebook’s “Oversight Board:” Move Fast with Stable Infrastructure and Humility’ (2019) 21 North Carolina Journal of Law & Technology 1
  - ‘Governing Online Speech: From “Posts-As-Trumps” to Proportionality and Probability’ [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3679607> accessed 1 December 2021
  - ‘What Kind of Oversight Board Have You Given Us?’ (2020) 2020 University of Chicago Law Review Online 1

- Duguay S, Burgess J and Suzor N, ‘Queer Women’s Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine’ (2020) 26 Convergence 237
- Dutkiewicz L and Krack N,
  - ‘All Eyes Riveted on the Trilogue Closed Doors of the Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online [Part I]’ (*CITIP blog*, 24 November 2020) <https://www.law.kuleuven.be/citip/blog/all-eyes-riveted-on-the-trilogue-closed-doors-of-the-proposal-for-a-regulation-on-preventing-the-dissemination-of-terrorist-content-online-part-i/> accessed 16 November 2022
  - ‘How to Notice without Looking: The “algorithmization” of Terrorist Content Moderation in the Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online [Part II] - CITIP Blog’ <https://www.law.kuleuven.be/citip/blog/how-to-notice-without-looking-the-algorithmization-of-terrorist-content-moderation-in-the-proposal-for-a-regulation-on-preventing-the-dissemination-of-terrorist-content-online-part-ii/> accessed 16 November 2022
- Dwoskin E and Zakrzewski C, ‘Facebook’s Independent Oversight Board Demands Transparency on Exemptions for Politicians’ *Washington Post* (21 September 2021) <https://www.washingtonpost.com/technology/2021/09/21/facebook-oversight-board-transparency-xcheck/> accessed 18 January 2023
- ‘EDPB-EDPS Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council Laying down Rules to Prevent and Combat Child Sexual Abuse | European Data Protection Board’ <https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-042022-proposal_en> accessed 20 January 2023
- European Digital Rights (EDRi),
  - ‘Coalition of Human Rights and Journalist Organisations Express Concerns for Free Speech’ (*European Digital Rights (EDRi)*) <https://edri.org/our-work/coalition-humn-rights-media-organisations-express-gave-concerns-free-speech/> accessed 9 February 2023
  - ‘French Avia Law Declared Unconstitutional: What Does This Teach Us at EU Level?’ (European Digital Rights (EDRi), 24 June 2020) <https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/> accessed 17 March 2023

- Europol,
  - ‘EU IRU Transparency Report 2019’ (*Europol*) <https://www.europol.europa.eu/media-press/newsroom/news/eu-iru-transparency-report-2019> accessed 10 February 2023
  - ‘Exploiting Isolation: Sexual Predators Increasingly Targeting Children during COVID Pandemic’ (*Europol*) <https://www.europol.europa.eu/media-press/newsroom/news/exploiting-isolation-sexual-predators-increasingly-targeting-children-during-covid-pandemic> accessed 24 January 2023
- Farokhmanesh M., ‘White Supremacists Who Used Discord to Plan Charlottesville Rally May Soon Lose Their Anonymity’ (The Verge, August 7, 2018) <https://www.theverge.com/2018/8/7/17660308/white-supremacists-charlottesville-rally-discord-plan> accessed February 28, 2023.
- Flew T, Martin F and Suzor N, ‘Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance’ (2019) 10 Journal of Digital Media & Policy 33
- ‘Fury over Facebook “Napalm Girl” Censorship’ *BBC News* (9 September 2016) <https://www.bbc.com/news/technology-37318031> accessed 2 December 2021
- Geese A, ‘Why the DSA could save us from the rise of authoritarian regimes’ [2022] Verfassungsblog <https://verfassungsblog.de/dsa-authoritarianism/> accessed 8 March 2023
- Geiger C. and Jütte BJ, ‘Platform Liability under Art. 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match’ (2021) 70 GRUR International 517.
- Gellert R and Wolters P, ‘The Revision of the European Framework for the Liability and Responsibilities of Hosting Service Providers’
- Ghosh D, ‘Facebook’s Oversight Board Is Not Enough’ [2019] *Harvard Business Review* <https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough> accessed 16 November 2022
- Giglio F, ‘The New Regulation on Addressing the Dissemination of Terrorist Content Online: A Missed Opportunity to Balance Counter-Terrorism and Fundamental Rights?’ (*CITIP blog*, 14 September 2021) <https://www.law.kuleuven.be/citip/blog/the-new-regulation-on-addressing-the-dissemination-of-terrorist-content-online/> accessed 3 February 2023
- Gilbert SA, ‘“I Run the World’s Largest Historical Outreach Project and It’s on a Cesspool of a Website.” Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians’ (2020) 4 Proceedings of the ACM on Human-Computer Interaction 19:1
- Gillespie T, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) <http://www.degruyter.com/document/doi/10.12987/9780300235029/html> accessed 1 December 2021

- Gorwa R, 'What Is Platform Governance?' (2019) 22 Information, Communication & Society 854
- Gorwa R, Binns R and Katzenbach C, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 205395171989794
- 'Governance | Oversight Board' <https://www.oversightboard.com/governance/> accessed 18 January 2023
- Grimmelmann J, 'The Virtues of Moderation' (LawArXiv 2017) preprint <https://osf.io/qwxf5> accessed 1 December 2021
- Grotius H, *De Jure Belli Ac Pacis. Libri Tres*
- 'Guidance Note on Content Moderation' (*Freedom of Expression*) <https://www.coe.int/en/web/freedom-expression/news/-/asset_publisher/thFVuWFiT2Lk/content/guidance-note-on-content-moderation> accessed 3 December 2021
- Guzel S., 'Article 17 of the CDSM Directive and the Fundamental Rights: Shaping the Future of the Internet,'
- Halfaker A. and R. Stuart Geiger RS, "Ores: Lowering Barriers with Participatory Machine Learning in Wikipedia" (2020) 4 Proceedings of the ACM on Human-Computer Interaction 1.
- Helberger N. and others, 'The EU's regulatory push against disinformation: What happens if platforms refuse to cooperate?,' (VerfBlog, 2022/8/05), <https://verfassungsblog.de/voluntary-disinfo/>, accessed February 20, 2023.
- 'Heller - Combating Terrorist-Related Content Through AI and.Pdf' <https://www.ivir.nl/publicaties/download/Hash_sharing_Heller_April_2019.pdf> accessed 2 December 2021
- Holt J, 'Cloud Policy: Anatomy of a Regulatory Crisis' (October 2017)
- Hummel K, 'The Christchurch Attacks: Livestream Terror in the Viral Video Age' (*Combating Terrorism Center at West Point*, 18 July 2019) <https://ctc.westpoint.edu/christchurch-attacks-livestream-terror-viral-video-age/> accessed 9 February 2023
- Husovec M, 'Will the DSA work?: On money and effort' [2022] Verfassungsblog <https://verfassungsblog.de/dsa-money-effort/> accessed 17 February 2023
- Jiang JA and others, 'Moderation Challenges in Voice-Based Online Communities on Discord' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1.
- Jiang JA, 'Toward a Multi-Stakeholder Perspective for Improving Online Content Moderation (Partial PhD in Philosophy)' (Department of Information Science, Faculty of the Graduate School of the University of Colorado 2020)
- Kabelka L,
  - 'EU Assessment of Child Abuse Detection Tools Based on Industry Data' (*www.euractiv.com*, 5 October 2022) <https://www.euractiv.com/section/digital/news/eu-assessment-of-child-abuse-detection-tools-based-on-industry-data/> accessed 25 January 202

- Killeen M, ‘EU Council Discusses Cross-Border Removal Orders to Fight Child Pornography’ (*www.euractiv.com*, 21 November 2022) <https://www.euractiv.com/section/digital/news/eu-council-discusses-cross-border-removal-orders-to-fight-child-pornography/> accessed 25 January 2023
- ‘MEPs Sceptical on EU Proposal to Fight Online Child Sexual Abuse’ (*www.euractiv.com*, 11 October 2022) <https://www.euractiv.com/section/digital/news/meps-sceptical-on-eu-proposal-to-fight-online-child-sexual-abuse/> accessed 25 January 2023

○ Komaitis K,

- ‘Can Mastodon Survive Europe’s Digital Services Act?’ (*Tech Policy Press*, 16 November 2022) <https://techpolicy.press/can-mastodon-survive-europes-digital-services-act/> accessed 26 January 2023
- ‘Infrastructure And Content Moderation: Challenges And Opportunities’ (*Techdirt*, 4 October 2021) <https://www.techdirt.com/2021/10/04/infrastructure-content-moderation-challenges-opportunities/> accessed 5 March 2023

○ Krack N and others, ‘AI in the Belgian Media Landscape. When Fundamental Risks Meet Regulatory Complexities’, *Artificial Intelligence and the Law*, vol 13 (Second Revised Edition, Jan De Bruyne and Cedric Vanleenhove (eds), Intersentia 2023) https://intersentia.com/en/artificial-intelligence-and-the-law-2nd-edition.html

○ Krack N,

- ‘Could Do Better! The European Commission’s Assessment of the EU Code of Practice on Disinformation Is out.’ (*KU Leuven Centre for IT and IP law*, 20 October 2020) <https://www.law.kuleuven.be/citip/blog/could-do-better-the-european-commissions-assessment-of-the-eu-code-of-practice-on-disinformation-is-out/> accessed 20 March 2023
- 'DSA Proposal and Disinformation - Should “Traditional Media” Be Exempted from Platform Content Moderation?’ (KU Leuven Centre for IT and IP law, 7 December 2021) <https://www.law.kuleuven.be/citip/blog/dsa-proposal-and-disinformation-should-traditional-media-be-exempted-from-platform-content-moderation/> accessed 20 March 2023.
- ‘MediaFutures Contributes in the Fight against Disinformation | Media Futures’ (*MediaFutures*, 20 June 2022) <https://mediafutures.eu/mediafutures-contributes-in-the-fight-against-disinformation/> accessed 20 March 2023

○ Kuczerawy A,

- ‘Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative’ (2015) 31 Computer Law & Security Review 46
- Intermediary Liability and Freedom of Expression in the EU: From Concepts to Safeguards (Intersentia 2018).
- ‘Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?’, *Disinformation and digital media as a challenge for democracy*, vol 6 (Cambridge 2020)

    - ‘Remedying Overremoval: The Three-Tiered Approach of the DSA’ [2022] Verfassungsblog <https://verfassungsblog.de/remedying-overremoval/> accessed 8 March 2023
    - ‘Safeguards for Freedom of Expression in the Era of Online Gatekeeping’ (20180914) 2017 Auteurs en Media 292
    - ‘Social Media Councils under the DSA: a path to individual error correction at scale?’, in: M. Kettemann (ed.), Platform://Democracy Project - Research Clinic Europe, commissioned by the Stiftung Mercator, and it is carried out by the Leibniz Institute for Media Research | Hans-Bredow-Institut (HBI) with support from the Humboldt Institute for Internet and Society (Berlin) and the Department of Theory and Future of Law of the University of Innsbruck (Austria). See more information https://leibniz-hbi.de/en/news/platform-councils-as-tools-to-democratize-hybrid-online-orders, 2023, forthcoming.
- Kuczerawy A and Dutkiewicz L, ‘Accessing Information about Abortion’ (Verfassungsblog, 28 July 2022) <https://verfassungsblog.de/accessing-information-about-abortion/> accessed 7 March 2023
- Kuklis L, ‘Media Regulation at a Distance: Video-Sharing Platforms in AVMS Directive and the Future of Content Regulation’
- Kwoka J and Valletti TM, ‘Scrambled Eggs and Paralyzed Policy: Breaking Up Consummated Mergers and Dominant Firms’ <https://papers.ssrn.com/abstract=3736613> accessed 1 March 2023
- Laidlaw EB, ‘A Framework for Identifying Internet Information Gatekeepers’ (2010) 24 International Review of Law, Computers & Technology 263
- Langvardt K, ‘Regulating Online Content Moderation’ (2018) 106 The Georgetown Law Journal <https://www.law.georgetown.edu/georgetown-law-journal/in-print/volume-106/volume-106-issue-5-june-2018/regulating-online-content-moderation/> accessed 10 February 2023
- Lawson EC, ‘New Research Shows Metaverse Is Not Safe for Kids’ (*Center for Countering Digital Hate | CCDH*, 30 December 2021) <https://counterhate.com/blog/new-research-shows-metaverse-is-not-safe-for-kids/> accessed 26 January 2023
- Lee H-E and others, ‘Detecting Child Sexual Abuse Material: A Comprehensive Survey’ (2020) 34 Forensic Science International: Digital Investigation 301022
- Leerssen P, ‘An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation’ (2023) 48 Computer Law & Security Review 105790
- ‘Lessons from Social Media for Creating a Safe Metaverse’ <https://itif.org/publications/2022/04/28/lessons-social-media-creating-safe-metaverse/> accessed 26 January 2023
- Lindroos A, ‘Addressing Norm Conflicts in a Fragmented Legal System: The Doctrine of Lex Specialis’ (2005) 74 Nordic Journal of International Law 27

- Llansó EJ and others, ‘Artificial Intelligence, Content Moderation, and Freedom of Expression’ (2020)
- Masnick M., ‘How Google's ContentID System Fails at Fair Use & the Public Domain’ (Techdirt, January 1, 2021) <https://www.techdirt.com/2012/08/08/how-googles-contentid-system-fails-fair-use-public-domain/> accessed February 20, 2023.
- ‘Mastodon Server Covenant for Joinmastodon.Org’ <https://joinmastodon.org/covenant> accessed 26 January 2023
- Matias JN, ‘The Civic Labor of Volunteer Moderators Online’ (2019) 5 Social Media + Society 2056305119836778
- ‘Meet Your Newest Community Moderator: AutoMod Is Here’(Discord Blog, October 31, 2022) <https://discord.com/blog/automod-launch-automatic-community-moderation> accessed February 25, 2023.
- ‘Meet the New ‘verse, Same as the Old ‘verse: Moderating the “Metaverse”’ (*Georgetown Law Technology Review*, 2 May 2022) <https://georgetownlawtechreview.org/meet-the-new-verse-same-as-the-old-verse-moderating-the-metaverse/GLTR-05-2022/> accessed 26 January 2023
- Michèle Finck, ‘Artificial Intelligence and Online Hate Speech, Centre on Regulation in Europe (CERRE), (2019).
- Miles T, ‘U.N. Investigators Cite Facebook Role in Myanmar Crisis’ Reuters (12 March 2018) <https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN> accessed 17 March 2023
- ‘Moderating the Fediverse: Content Moderation on Distributed Social Media by Alan Z. Rozenshtein :: SSRN’ <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213674> accessed 24 January 2023
- ‘OHCHR | Report of the Special Rapporteur to the General Assembly on AI and Its Impact on Freedom of Opinion and Expression’ <https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx> accessed 3 December 2021
- ‘Opinion Piece: The DSA Also Works “in the Metaverse” – If It Is Enforced Well’ (14 December 2022) <https://www.stiftung-nv.de/en/publication/opinion-piece-dsa-also-works-metaverse-if-it-enforced-well> accessed 26 January 2023
- Oruç TH, ‘The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in Light of Recent Developments: Is It Still Necessary to Maintain It?’ (2022) 13 JIPITEC <http://www.jipitec.eu/issues/jipitec-13-3-2022/5555>
- Oversight Board
    - ‘Oversight Board Publishes First Annual Report | Oversight Board’ <https://www.oversightboard.com/news/322324590080612-oversight-board-publishes-first-annual-report/> accessed 19 January 2023

  - ‘Oversight Board Announces Plans to Review More Cases, and Appoints a New Board Member’ <https://www.oversightboard.com/news/943702317007222-oversight-board-announces-plans-to-review-more-cases-and-appoints-a-new-board-member/> accessed 20 March 2023
  - ‘Oversight Board Cases’ (Meta Transparency Centre) <https://transparency.fb.com/oversight/oversight-board-cases/> accessed 20 March 2023

- ‘Oxford Word of the Year 2016’ (Oxford Languages), available at: <https://languages.oup.com/word-of-the-year/2016/> accessed February 20, 2023.
- Parks L, ‘Dirty Data: Content Moderation, Regulatory Outsourcing, and The Cleaners’ (2019) 73 Film Quarterly 11
- PBS, ‘The Cleaners’, <https://www.pbs.org/independentlens/documentaries/the-cleaners/>, accessed 20 January 2023.
- Peters J., ‘Discord Bans pro-Trump Server 'the Donald’’ (The Verge, January 9, 2021) <https://www.theverge.com/2021/1/8/22221579/discord-bans-the-donald-server-reddit-subreddit> accessed February 26, 2023.
- ‘Photographer Nick Ut: The Napalm Girl | Buy Photos | AP Images | Collections’ <http://www.apimages.com/Collection/Landing/Photographer-Nick-Ut-The-Napalm-Girl-/ebfc0a860aa946ba9e77eb786d46207e> accessed 2 December 2021
- Pickup EL, ‘The Oversight Board’s Dormant Power to Review Facebook’s Algorithms’ <https://openyls.law.yale.edu/handle/20.500.13051/18219> accessed 19 January 2023
- Pirkova E, ‘The EU Digital Services Act Won’t Work without Strong Enforcement’ (*Access Now*, 9 December 2021) <https://www.accessnow.org/eu-dsa-enforcement/> accessed 8 March 2023
- ‘Position of the EU DisinfoLab on the 2022 Code of Practice on Disinformation’ (EU DisinfoLab, September 8, 2022), <https://www.disinfo.eu/advocacy/eu-disinfolabs-position-on-the-2022-code-of-practice-on-disinformation/> accessed February 20, 2023.
- Quintais JP and Schwemer SF, ‘The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?’ (2022) 13 European Journal of Risk Regulation 191
- Quintais JP, ‘Between Filters and Fundamental Rights: How the Court of Justice saved Article 17 in C-401/19 - Poland v. Parliament and Council,’ (VerfBlog, 2022/5/16), <https://verfassungsblog.de/filters-poland/>
- Raji ID and others, ‘Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing’ [2020] arXiv:2001.00964 [cs] <http://arxiv.org/abs/2001.00964> accessed 27 July 2021

- Rauchegger C and Kuczerawy A, ‘Injunctions to Remove Illegal Online Content under the Ecommerce Directive: Glawischnig-Piesczek’ <https://papers.ssrn.com/abstract=3728597> accessed 25 November 2022
- ‘Regulating Content Moderation in Europe beyond the AVMSD’ (*Media@LSE*, 25 February 2020) <https://blogs.lse.ac.uk/medialse/2020/02/25/regulating-content-moderation-in-europe-beyond-the-avmsd/> accessed 28 February 2023
- Roberts ST, *Behind the Screen* (Yale University Press 2021) <https://yalebooks.yale.edu/9780300261479/behind-the-screen> accessed 10 February 2023
- Roose K., ‘This Was the Alt-Right's Favorite Chat App. Then Came Charlottesville’ (The New York Times, August 15, 2017) <https://www.nytimes.com/2017/08/15/technology/discord-chat-app-alt-right.html> accessed February 25, 2023.
- ‘Rules Enforcement - Twitter Transparency Center’ <https://transparency.twitter.com/en/reports/rules-enforcement.html> accessed 1 December 2021
- ‘Sama Exploit Facebook Moderators and Call It “Ethical”. Help Us Stop Them’ (*Foxglove*) <https://www.foxglove.org.uk/campaigns/sama-bcorp/> accessed 23 February 2023
- Sartor G and Loreggia A, ‘Study for the European Parliament on the Impact of Algorithms for Online Content Filtering or Moderation’ (2020) <https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)657101> accessed 26 January 2023
- Satariano A and Isaac M, ‘The Silent Partner Cleaning Up Facebook for $500 Million a Year’ *The New York Times* (31 August 2021) <https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html> accessed 22 February 2023
- Sato Y., ‘Non-English Wikipedia is rife with misinformation. Here’s how to fix it’ (Fast Company, 27 July 2021) <https://www.fastcompany.com/90666412/non-english-wikipedia-misinformation accessed> accessed March 17, 2023.
- Seering J and others, ‘Moderator Engagement and Community Development in the Age of Algorithms’ (2019) 21 New Media & Society 1417
- ‘Self-Regulation and “Hate Speech” on Social Media Platforms’ (*ARTICLE 19*, 2 March 2018) <https://www.article19.org/resources/self-regulation-hate-speech-social-media-platforms/> accessed 21 February 2023
- ‘Social Media Councils’ (*ARTICLE 19*) <https://www.article19.org/social-media-councils/> accessed 21 February 2023
- Somers C, ‘The Proposed CSAM Regulation: Trampling Privacy in the Fight against Child Sexual Abuse?’ (*CITIP blog*, 3 January 2023) <https://www.law.kuleuven.be/citip/blog/the-proposed-csam-regulation-trampling-privacy-in-the-fight-against-child-sexual-abuse/> accessed 20 January 2023

- Spotify, ‘Spotify Continues to Ramp Up Platform Safety Efforts with Acquisition of Kinzen’ (*Spotify*, 5 October 2022) <https://newsroom.spotify.com/2022-10-05/spotify-continues-to-ramp-up-platform-safety-efforts-with-acquisition-of-kinzen/> accessed 22 February 2023
- Steiger M and others, ‘The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support’, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3411764.3445092> accessed 20 January 2023
- Sun H and Ni W, ‘Design and Application of an AI-Based Text Content Moderation System’ (2022) 2022 Scientific Programming e2576535
- ‘The Facebook Files’ *Wall Street Journal* (1 October 2021) <https://www.wsj.com/articles/the-facebook-files-11631713039> accessed 18 January 2023
- ‘The Metaverse Has a Groping Problem Already | MIT Technology Review’ <https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/> accessed 7 March 2023
- ‘The online regulation series| European Union (Update) - Tech Against Terrorism’ (10 December 2021) <https://www.techagainstterrorism.org/2021/12/10/the-online-regulation-series-european-union-update/, https://www.techagainstterrorism.org/2021/12/10/the-online-regulation-series-european-union-update/> accessed 7 March 2023
- ‘The Underwear Rule - Children’s Rights - Publi.Coe.Int’ (*Children’s Rights*) <https://www.coe.int/en/web/children/underwear-rule> accessed 20 January 2023
- Tworek H, ‘Social Media Councils’, *Models for Platform Governance* (2019)
- University of Standford, Global Digital Policy Incubator, Cyper Policy Center, ‘Social Media Councils: From Concept to Reality - Conference Report’ <https://cyber.fsi.stanford.edu/gdpi/content/social-media-councils-concept-reality-conference-report> accessed 21 February 2023
- ‘YouTube Community Guidelines Enforcement – Google Transparency Report’ <https://transparencyreport.google.com/youtube-policy/removals?hl=en> accessed 1 December 2021
- Wagner B. and others, ‘Reimagining Content Moderation and Safeguarding Fundamental Rights’ (May 2021), <https://www.greens-efa.eu/files/assets/docs/alternative_content_web.pdf> accessed February 22, 2023.
- ‘What Is Content Moderation?’ (*Trust and Safety Professional Association*) <https://www.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/what-is-content-moderation/> accessed 14 February 2023
- ‘Who Gets To Be 'Notable' — And Who Doesn't? Gender Bias On Wiki’ (National Public Radio, 13 July 2021) <https://www.npr.org/2021/07/13/1015754856/who-gets-to-be-notable-and-who-doesnt-gender-bias-on-wiki> accessed March 17, 2023.

- ‘Why Facebook Is Losing the War on Hate Speech in Myanmar’ *Reuters* (15 August 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 23 February 2023
- Wikimedia Foundation,
  - ‘Terms of Use’ (Wikimedia Foundation Governance Wiki) <https://foundation.wikimedia.org/wiki/Terms_of_Use/en> accessed March 6, 2023.
  - ‘EU DSA USERBASE Statistics’ (Wikimedia Foundation Governance Wiki) <https://foundation.wikimedia.org/wiki/Legal:EU_DSA_Userbase_Statistics> accessed March 6, 2023.
  - ‘Ores - MediaWiki’ (Powered by MediaWiki, June 8, 2022) <https://www.mediawiki.org/wiki/ORES> accessed February 20, 2023.
  - ‘Administration’ (Wikipedia, November 24, 2022) <https://en.wikipedia.org/wiki/Wikipedia:Administration> accessed March 6, 2023
  - ‘Copyrights’ (Wikipedia, December 20, 2022) <https://en.wikipedia.org/wiki/Wikipedia:Copyrights> accessed February 24, 2023.
  - ‘Non-Free Content’ (Wikipedia, January 15, 2023) <https://en.wikipedia.org/wiki/Wikipedia:Non-free_content> accessed February 24, 2023.
  - ‘Our Work’ (Wikimedia Foundation, January 19, 2023) <https://wikimediafoundation.org/our-work/> accessed March 6, 2023.
  - ‘Protection Policy’ (Wikipedia, February 24, 2023) <https://en.wikipedia.org/wiki/Wikipedia:Protection_policy> accessed March 6, 2023.
  - ‘What Wikipedia Is Not’ (Wikipedia, March 4, 2023) <https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#Wikipedia_is_not_censored> accessed March 6, 2023.
- Wilman F, ‘THE DIGITAL SERVICES ACT (DSA): AN OVERVIEW’
- Wong D and Floridi L, ‘Meta’s Oversight Board: A Review and Critical Assessment’ [2022] Minds and Machines <https://doi.org/10.1007/s11023-022-09613-x> accessed 21 December 2022
- Yildirim EO and others, Freedom to Share: How the Law of Platform Liability Impacts Licensors and Users,’ (Creative Commons Medium, 2021), <https://medium.com/creative-commons-we-like-to-share/freedom-to-share-how-the-law-of-platform-liability-impacts-licensors-and-users-84d86adade4e>
- Yildirim EO, ‘Silenced, Chilled, and Jailed: The New Turkish Law Criminalizes Disseminating ‘Disinformation,’ (VerfBlog, 20 October 2022), <https://verfassungsblog.de/silenced-chilled-and-jailed/>

Court-cases

- *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECJ Case C-360/10
- *Erbakan v Turkey [2006] ECtHR 59405/00*
- *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECJ Case C-18/18
- *Google France SARL and Google Inc v Louis Vuitton Malletier SA (C-236/08),*
- *Google France SARL v Centre national de recherche en relations humaines (CNRRH) SARL and Others (C-238/08)* [2010] ECJ Joined cases C-236/08 to C-238/08
- *Google France SARL v Viaticum SA and Luteciel SARL (C-237/08)*
- *Handyside v the United Kingdom [1976] ECtHR 5493/72*
- *Joined Cases C-682/18 and C-683/18: Judgment of the Court (Grand Chamber) of 22 June 2021 (requests for a preliminary ruling from the Bundesgerichtshof — Germany) — Frank Peterson v Google LLC, YouTube LLC, YouTube Inc, Google Germany GmbH (C-682/18) and Elsevier Inc v Cyando AG (C-683/18)* (ECJ)
- Judgment of the Court (Third Chamber), 16 February 2012 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV*, EU:C:2012:85
- *L'Oréal SA et autres contre eBay International AG et autres* [2011] Cour de justice Affaire C-324/09
- Lenz v. Universal Music Corp., 801 F.3d 1126 (US Court of Appeals, 9th Cir. 2015)
- *Republic of Poland v European Parliament and Council of the European Union* [2022] ECJ Case C-401/19
- *Seurot v La France (dec) [2004] ECtHR 57383/00*

EU legislative and policy documents

- Communication from the Commission (...)on the EU Security Union Strategy 2020 [COM(2020) 605 final]
- Council conclusions on shaping Europe's digital future Brussels, 9 June 2020 (OR. en) 8711/20
- Council Decision of 29 May 2000 to combat child pornography on the Internet 2000 (OJ L)
- Council framework Decision 2004/68/JHA of 22 December 2003 on combating the sexual exploitation of children and child pornography 2003 (OJ L)
- Decision No 854/2005/EC of the European Parliament and of the Council of 11 May 2005 establishing a multiannual Community Programme on promoting safer use of the Internet and new online technologies (Text with EEA relevance) 2005 (OJ L)
- Directive (EU) 2001/29 on Copyright and Information Society
- Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities, OJ L 303, 28.11.2018, p. 69–92

- Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications Code (Recast) (Text with EEA relevance)Text with EEA relevance 2018
- Directive (EU) 2019/790 on Copyright in the Digital Single Market
- Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) 2009
- Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA 2011
- European Commission,
    - ‘A multi-dimensional approach to disinformation, Report of the independent High level Group on fake news and online disinformation’, (European Commission, 12 March 2018).
    - ‘Code of Practice on Disinformation: New Transparency Centre Provides Insights and Data on Online Disinformation for the First Time,’ (European Commission, 9 February 2023) <https://ec.europa.eu/commission/presscorner/detail/en/mex_23_723> accessed March 6, 2023.
    - ‘Disinformation: Commission Welcomes the New Stronger and More Comprehensive Code of Practice on Disinformation’ (*European Commission*, 16 June 2022) <https://ec.europa.eu/commission/presscorner/detail/en/IP_22_3664> accessed 20 March 2023
    - ‘EU Centre to Prevent and Combat Child Sexual Abuse. Why Oblige Platforms to Detect, Report and Remove Online Child Sexual Abuse’ <https://home-affairs.ec.europa.eu/whats-new/campaigns/legislation-prevent-and-combat-child-sexual-abuse/eu-centre-prevent-and-combat-child-sexual-abuse_en> accessed 25 January 2023
    - ‘EU Internet Forum Committed to an EU-Wide Crisis Protocol’ (European Commission) <https://ec.europa.eu/commission/presscorner/detail/en/IP_19_6009> accessed 17 March 2023.
    - ‘EU Strategy for a More Effective Fight against Child Sexual Abuse’ <https://home-affairs.ec.europa.eu/policies/internal-security/child-sexual-abuse/eu-strategy-more-effective-fight-against-child-sexual-abuse_en> accessed 25 January 2023
    - ‘European Union Internet Forum (EUIF)’ <https://home-affairs.ec.europa.eu/networks/european-union-internet-forum-euif_en> accessed 9 February 2023
    - Europol to Boost the EU’s Resilience’ (*European Commission - European Commission*, 9 December 2020)

<https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2326> accessed 10 February 2023

  - ‘Feedback and Statistics: Proposal for a Regulation. Fighting Child Sexual Abuse: Detection, Removal and Reporting of Illegal Content Online’ <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12726-Fighting-child-sexual-abuse-detection-removal-and-reporting-of-illegal-content-online/feedback_en?p_id=30786148> accessed 25 January 2023
  - ‘Fighting Child Sexual Abuse’ (*European Commission - European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/IP_22_2976> accessed 20 January 2023
  - ‘Fighting child sexual abuse of children: Commission proposes new rules to protect children’ (*European Commission*) <https://ec.europa.eu/commission/presscorner/detail/es/ip_20_2463> accessed 20 January 2023
  - ‘REFIT – Making EU Law Simpler, Less Costly and Future Proof’ <https://commission.europa.eu/law/law-making-process/evaluating-and-improving-existing-laws/refit-making-eu-law-simpler-less-costly-and-future-proof_en> accessed 17 March 2023.
  - ‘Security Union: A Counter-Terrorism Agenda and Stronger European Commission, /* COM/2011/0942 final - 2012/ () */ COMMISSION COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A coherent framework for building trust in the Digital Single Market for e-commerce and online services.

- ‘European Parliament Resolution of 11 March 2015 on Child Sexual Abuse Online (2015/2564(RSP)’ <https://www.europarl.europa.eu/doceo/document/TA-8-2015-0070_EN.html> accessed 23 January 2023
- European Council, ‘The EU’s Response to Terrorism’ (15 December 2022) <https://www.consilium.europa.eu/en/policies/fight-against-terrorism/> accessed 3 February 2023
- European Parliament resolution of 14 December 2017 on the implementation of Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography (2015/2129(INI)) 2017
- P9_TA(2021)0319 Use of technologies for the processing of data for the purpose of combating online child sexual abuse (temporary derogation from Directive 2002/58/EC) ***I European Parliament legislative resolution of 6 July 2021 on the proposal for a regulation of the European Parliament and of the Council on a temporary derogation from certain provisions of Directive 2002/58/EC of the European Parliament and of the Council as regards as the use of technologies by

number-independent interpersonal communications service providers for the processing of personal and other data for the purpose of combatting child sexual abuse online (COM(2020)0568 — C9-0288/2020 — 2020/0259(COD)) P9_TC1-COD(2020)0259 Position of the European Parliament adopted at first reading on 6 July 2021 with a view to the adoption of

- Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down rules to prevent and combat child sexual abuse 2022
- Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a temporary derogation from certain provisions of Directive 2002/58/EC of the European Parliament and of the Council as regards the use of technologies by number-independent interpersonal communications service providers for the processing of personal and other data for the purpose of combatting child sexual abuse online 2020
- Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online A contribution from the European Commission to the Leaders' meeting in Salzburg on 19-20 September 2018 2018 [COM/2018/640 final]
- Regulation (EU) 2016/679 of the European Parliament and of the Council - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation) 2016 [Regulation (EU) 2016/679] 88
- Regulation (EU) 2021/… of the European Parliament and of the Council on a temporary derogation from certain provisions of Directive 2002/58/EC as regards the use of technologies by providers of number-independent interpersonal communications services for the processing of personal and other data for the purpose of combating online child sexual abuse 2021
- Regulation (EU) 2021/1232 of the European Parliament and of the Council of 14 July 2021 on a temporary derogation from certain provisions of Directive 2002/58/EC as regards the use of technologies by providers of number-independent interpersonal communications services for the processing of personal and other data for the purpose of combating online child sexual abuse (Text with EEA relevance) 2021 (OJ L)
- Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance) 2021 (OJ L)
- Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) 2022 (OJ L)Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) 2022 (OJ L)

AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu