AI4media
ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

www.ai4media.eu

@ai4mediaproject

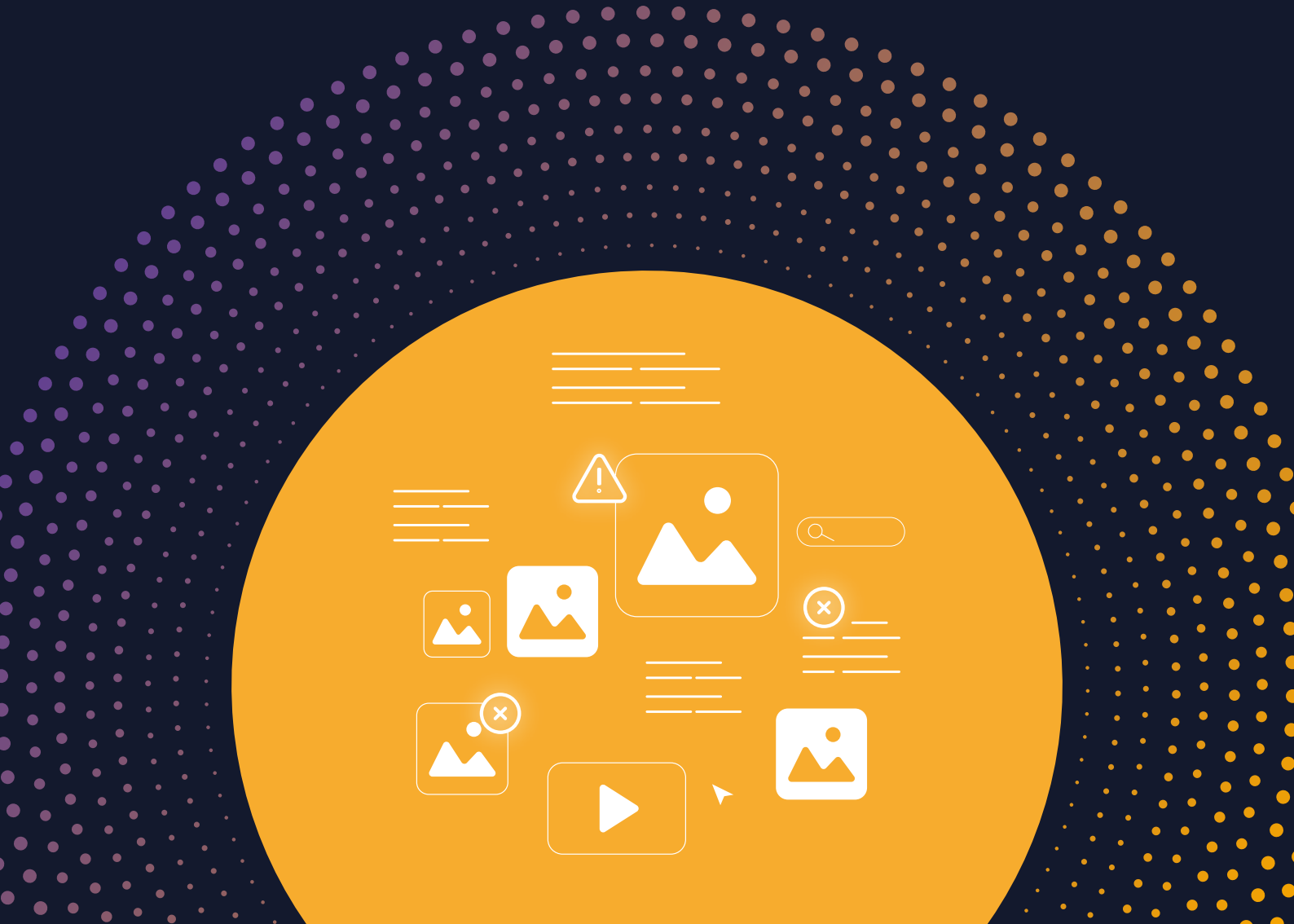
AI4Media Results in Brief: **Report on Policy for Content Moderation**

Authors:
Noémie Krack, Lidia Dutkiewicz and Emine Ozge Yildirim (KU Leuven)

This factsheet is a summary of the Report on Policy for Content Moderation developed by the AI4Media project. It presents an overview of the EU policy initiatives on content moderation as well as alternative approaches to content moderation by online platforms and civil society.

The present factsheet will keep the structure of the deliverable and refer to the pages of the deliverable for further information on the related section.

The full deliverable "Report on Policy for Content Moderation" can be accessed through this **LINK**.



1

Introduction

The Internet and social media 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. It has become clear that some form of content moderation (e.g., content removal, accounts suspension) is necessary in order to make it safer for users and to tackle power asymmetries and unlimited platforms' power over what wesee online.

The EU legislative efforts have intensified in the last few years, as the possible threats of not moderating content online became more obvious nowadays and to ensure the creation of a safe and harmonised EU digital single market.

Several approaches envisaged for content moderation include self-regulation, coregulation, 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, ... The concerns over power imbalance, delegated regulation of speech from public authorities to private actors, the lack of legitimacy of private actors to set speech rules become a topic of public debate. What these debates show is that content moderation regulation is a complex balancing exercise.

In the following, we provide a brief overview of the EU policy initiatives on content moderation as well as alternative approaches to content moderation by online platforms and civil society. We assess the challenges and advantages of these instruments and diverging approaches and outline policy recommendations for the future of content moderation in the EU. More specifically, the document 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.

For more information, we refer the reader to the full version of D6.2, Section 2, pp. 14-16.

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

2

Evolution of the EU Content Moderation Regulation

Content moderation: what is it and who does it involve?

Content moderation may be understood as the "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse."[1] 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). 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. 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.

The **content moderation value chain** involves different actors. Intermediary services (1) including hosting services providers (2) and their internal content moderation teams often composed of Trust and Safety teams with policy, operation and technical teams, content moderation sub-contractors (3), content moderators (4) and end-users flagging or monitoring content (5). Much is yet to be known about the human aspect of content moderation including the working conditions and contractual arrangements with human content moderators. Same goes for the **content moderation infrastructures**. It appears that in some cases the content moderation outsourcing took place in regions where the company did not have offices or language expertise leading to negative effects. In addition, the various layers of the internet are no longer distinguishable, and content moderation is composed of different interfaces and infrastructure layers[2]. More transparency on the role and responsibilities of infrastructure providers is necessary for having a complete picture of the content moderation landscape and challenges.

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

The **scale of content moderation** by platforms rose to unprecedented numbers. Mistakes in enforcing any rule are therefore inevitable: it will always be possible to find examples of both false positives (something is wrongly classified as objectionable) and false negatives (the automated tool misses something that should have been classified as objectionable)[3].

When it comes to 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. Legally required removals are shaped by content moderation legislations detailing what obligations are foreseen for what type of illegal content[4]. 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[5]. It's often now referred to as the platform's governance. Thanks to the freedom to conduct business they are free to decide in their terms what content can be hosted on their platform as long as it's not considered illegal by a legislation. 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.

Content moderation is a **powerful mechanism**. It is being analysed through a growing body of literature on platform governance. It analyses the moving power relations between the private actors, including internet and information technology (IT) companies, social media platforms, and public authorities but also how content moderation regulation can constitute a grip, a policy lever for a public authority to get some control of the increasingly powerful tech actors[6].

The section also devotes some focus on **AI systems used in content moderation efforts**.

Automated tools bring advantage in terms of scale, cost savings, and speedier decisions[7]. They also promise to relieve workers from the psychological trauma that comes with content moderation. 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, geo-blocking, and account takedown)"[8]. 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. Algorithmic content moderation involves a range of techniques from statistics and computer science but two main systems are: the matching/hashing and the predictive systems.

Challenges and limitations of algorithmic content moderation.

The use of AI systems in content moderation brings a set of challenges and limitations from a technical perspective and from a fundamental rights perspective.

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

This includes the **lack of contextual interpretation**. The AI content moderation tools are not yet able to understand context, irony, or satire[9]. This can lead to unjustified removals of legal uses of illegal material (such as for educational, artistic, journalistic, or research purposes, or awareness-raising purposes against the illegal activity). 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[10].

Second, there is a **lack of quality, diversity and inclusivity** in the data used. There is indeed currently a lack of representative, well-annotated datasets for machine learning training. For instance, local languages classifiers are missing and "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"[11].

In addition, defining in a specific context what constitutes harmful or illegal content is a **socio-political** matter and varies across countries and jurisdictions. This can lead to different opinions on the same content.

The use of AI systems may pose a challenge to all **fundamental rights**, but when it comes to content moderation, some are particularly at stake. Both false positives and false negatives impact the right to **freedom of expression**, including the freedom to impart and receive information but also indirectly create a chilling effect or prior restraints to free expression. The **right to privacy and the right to protection of personal data** will also be impacted as content moderation systems require the processing of a range of personal data which can include sensitive personal data. The right to **equality and non-discrimination** can also be harmed as algorithmic systems have the potential to reproduce and amplify existing biases. They can perform badly on data related to underrepresented groups, including racial and ethnic minorities, non-dominant languages, and/or political leanings[12]. This can lead to disadvantages such as censoring, preventive removal, unfavourable ranking, shadow banning, blacklisting of keywords by these communities. The **right to a fair trial and effective remedy** is also impacted as entrusting private stakeholders to take decisions on what is legal and what is illegal content puts a great deal of power in their hands without democratic control. 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[13].

For more information, we refer the reader to D6.2, Sections 3.1.3-3.1.4, pp. 27-33.

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

Content Moderation Landscape in Europe

The **EU regulatory framework on content moderation** is increasingly complex and has been differentiated over the years depending on the category of the online platform, the type of content, and the nature of the legal instrument (hard-law, soft-law, or self-regulation). 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 (VSPs). 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. 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. category of the online platform, the type of content, and the nature of the legal instrument (hard-law, soft-law, or self-regulation). 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 (VSPs). 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. 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[14].

For more information, we refer the reader to D6.2, Sections 3.2, pp. 33-78.

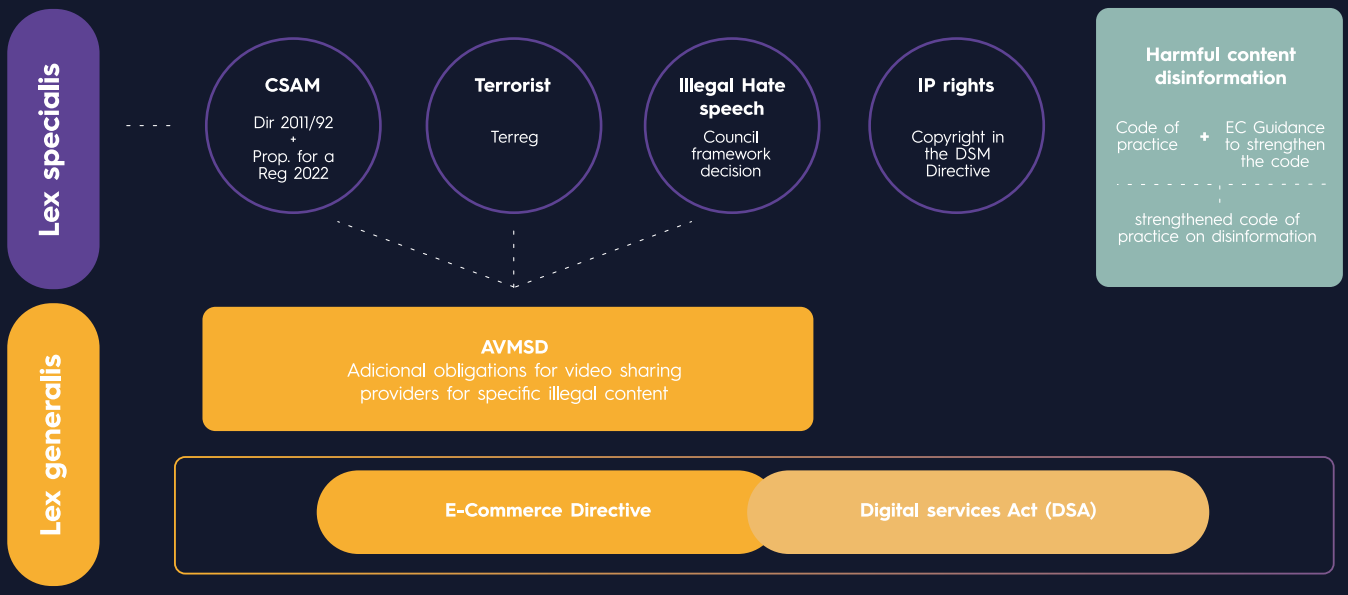


Figure 1 - Overview of the EU content moderation landscape[15]

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

Lex generalis instruments relevant to content moderation

Here, we outline the lex generalis instruments in the EU, namely the e-Commerce Directive, the Digital Services Act, and the Audio-visual Media Service Directive.

For more information, we refer the reader to D6.2, Section 3.2.1, pp. 34-51.

1

E-commerce Directive

The **e-commerce Directive**, 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. The Directive applies to any kind of illegal or infringing content. It provides for horizontal **liability exemptions** for the illegal content/goods/services present on the intermediary services posted or generated by third parties (users). 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. To benefit from the liability exemption hosting providers must 1) not have actual knowledge of illegal activity or information; 2) act expeditiously to remove or to disable access to the information upon obtaining such knowledge or awareness. The scope of hosting exemptions is quite broad as the case law of the Court of Justice of the European Union (CJEU) confirmed its applicability to marketplaces and social media. The Directive prohibits EU Member States to impose on intermediary service providers a **general obligation to monitor content** that they transmit or store. The prohibition of monitoring obligations does not concern monitoring obligations in a specific case. Criticism arises over the fragmented interpretation and legal uncertainty on certain e-commerce directive concepts. The lack of uniform rules for notice and action safeguards and procedures across the EU was also underlined as potentially leading to over-removal of content. The case-law also led to uncertainty about the use of AI tools to moderate content which caused conflicting interpretations of the prohibition of general monitoring obligation. The need for clear and harmonised evidence-based rules on responsibilities and accountability for digital services that would guarantee internet intermediaries and users an appropriate level of legal certainty was underlined and the revision of the e-commerce Directive started already in 2010. It led to the adoption of the DSA (see below)..

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

2 Audio-visual Media Service Directive

The **Audio-visual Media Service Directive** is the cornerstone of audio-visual media regulation in the EU. The text got revised in 2016 bringing major changes with regard to the broadening of the scope to include VSPs. The AVMSD is the first legal instrument that provides a catalogue of both procedural (e.g., complaint and redress mechanisms) and technical (e.g., age verification and parental control systems) measures which must be implemented by the VSPs. VSPs also have to protect minors from content that may impair their physical, mental or moral development. An oversight framework has been created to check where national authorities were given the responsibility of verifying that VSPs have adopted "appropriate measures" to deal with different types of content. As critics point out however, the AVMSD has a very narrow scope of application, namely only video content is covered and only to the extent that services are offered to the general public. The question of the applicability of the instrument on videos present on social media has been quite controversial. To solve this issue, the EC adopted Guidelines in 2020 but they are not legally binding and are open to various interpretations. In addition, the protection granted by AVSMD applies when the content is illegal because disseminating it constitutes a crime at the Union level, leaving the material scope to only very specific crimes. The AVMSD is a minimum harmonisation instrument laying down the minimum rules. This leaves the opportunity for the EU MS to go further but it also leads to divergent application of the AVMSD rules.

3 Digital Services Act

The **Digital Services Act** (DSA) 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. 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 category of actors are the following: intermediary services providers, hosting providers, online platforms and very large online platforms (VLOPs) and very large online search engines (VLOSEs). The regulation follows an asymmetric approach where a set of rules corresponds to one or more of these categories. The DSA maintains the liability rules for providers of intermediary services set out in the e-commerce Directive. The choice to maintain it was motivated by fundamental rights protection and legal certainty. The goal was to avoid a situation when platforms over remove content "just in case", in order to avoid potential liability if the content was illegal.

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

The DSA provides the first legal definition of content moderation. Content moderation is explicitly defined to include not just content removals (takedowns) or account suspension, but also demonetisation and visibility restrictions. The DSA contains many new provisions aimed at improving content moderation and better tackling illegal content disseminated through intermediary services. This include the clarification of the content and the scope of the national orders, the obligation to provide yearly transparency report on key elements of content moderation, the harmonisation of the notice and action framework to guarantee equal rights to end-users, the obligation to provide a statement of the reasons to be communicated to end-users following a content moderation decision. The DSA also sets three routes for redress against content moderation decisions giving end-users several options. It creates a category of trusted flaggers where the notice of these entities will be treated more rapidly. The VLOPs and VLOSEs will also have the obligation to self-assess and mitigate the systemic risks posed by their services including by content moderation practices. The text also integrates a crisis response mechanism granting some exceptional rights to the EC to request the adoption of urgent measures from private entities including on content moderation.

Many criticisms about the e-commerce Dir. got addressed the DSA but others remain. The issue of the accessibility of the information is not detailed and doubts remain about the enforcement. 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. What it means exactly is unclear. Much power is granted to the EC through the crisis response mechanisms, some fear that having a body unilaterally declaring an EU-wide state of emergency would enable far-reaching restrictions of freedom of expression. Some questions remain about the interaction of the lex specialis being a Regulation (directly applicable) and lex generalis (Directive). Strong enforcement will be key to materialise all the promises of this ambitious and necessary legislation. It remains to be seen if the text is future-proof and will cover services which go beyond social media platforms, such as metaverse.

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

Lex specialis and soft-law instruments applicable to Illegal Content and Harmful Content

Below, we outline 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.

For more information, we refer the reader to D6.2, Section 3.2.2, pp. 51-78.

1

Terrorist content

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. In May 2021, the European Commission adopted a Regulation on preventing the dissemination of terrorist content (TERREG). 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 time window 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. Of course, 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[16]. The TERREG enables MS restrictions on online speech after only a minimal review (not even judicial review needed) and sets a very short 1-hour window for action for intermediary services to act upon order receipt. This incentivises hosting providers to have a more proactive approach to avoid sanctions and rely on algorithmic moderation with all the downside already explained in Section 3.1. This creates risks for the right to freedom of expression and concerns about censorship. In addition, the lack of transparency for public-private security collaboration has been pointed out by some critics. Lastly, in 2020, the French Constitutional Court struck down the so-called Avia Law with similar provisions. Perhaps some TERREG provisions will be interpreted in the future by the CJEU in a similar way.

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

2 Copyright protected content

The Directive 2019/790/EC on Copyright in the Digital Single Market (CDSM) came into force in 2019. The Dir. sets out various provisions aiming to modernise the EU copyright framework in order to create a fairer marketplace for online content. The CDMS provisions transparency and balance in the contractual relations between content creators and producers and publishers. Its most infamous provision is art. 17 which has caused a lot of debate. It imposes direct liability on online content-sharing service providers (OCSSPs) for copyright-protected works or other protected subject-matter uploaded by users. It is justified by the fact that OCSSPs perform an act of communication to the public when they give the public access to copyright-protected content hence need to obtain authorisation or conclude a licensing agreement. Platforms could avoid liability in relation to user-generated content (UGC) infringing copyright in certain cases detailed in article 17.4 of the CDSM. The fear of liability may encourage some 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. This leads to removal of legitimate content. Poland filed an action for annulment of Art. 17 with the CJEU claiming that the Article violates freedom of expression (C-401/19). The request was dismissed as the Court found that the CSDMS provides adequate procedural safeguards and strikes a fair balance between different rights.

3 Child sexual abuse material

In 2011, the Child Sexual Abuse Material (CSAM) started to be regulated through EU legislation with the Child Sexual Abuse and Exploitation Directive (CSAED). The directive has set up minimum rules concerning the definition of criminal offences and sanctions in the area of child sexual exploitation and abuse. Since the expansion of the notion of electronic communication services in the European Electronic Communication Code (EECC), 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 Directive[17]. To fix this issue, the EC has adopted an interim CSAM regulation in July 2021 which will last until August 2024. In 2022, a proposal for a regulation laying down rules to prevent and combat child sexual abuse has been released to replace the interim regulation. 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. The proposal also creates a new independent

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

EU Centre on Child Sexual Abuse with several missions. The proposal for a regulation is now being debated and negotiated by EU policymakers (EP and Council). It provides an important shift in the content moderation regulation of CSAM from a voluntary practice to binding obligations on providers. However well intended, the current 2022 proposal has been subject to criticisms from scholars, EU co-legislators, and civil society. The criticism focuses on risks that the proposal's provision brings to the proportionality principle, data protection, and the right to privacy. The proposal introduces a general scanning obligation for messaging services which may lead to mass surveillance practices. As a result, an important debate in the EP has been started, and additional concerns were raised by the opinion of the EDPB-EDPS. Overall, 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.

4 Hate speech

In May 2016, the European Commission agreed with Facebook, Microsoft, Twitter and YouTube a Code of conduct on countering illegal hate speech online. The Code sets up a series of commitments encouraging platforms to: provide publicly available information on how to submit a notice flagging the hateful content; to 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; to 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. The Code has faced massive criticism, especially from the freedom-of-expression and digital rights civil society organisations. The implementation of the Code can lead to more censorship by private companies, and, therefore, have a chilling effect on freedom of expression. "Hateful content" is a vague term that could encompass mere vulgar abuse and there is a risk that platforms' understanding of these notions can go beyond, or even have no direct connection to the definitions established by the law[18]. In addition, there is a lack of transparency in the reporting systems of the Code and a lack of sufficient safeguards against misuse of the notice procedure. The same criticisms for TERREG can be expressed here when it comes to the 24 hours deadline for taking down illegal hate speech content.

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

5

Disinformation

Disinformation is not a uniformly defined concept, hence providing a legal definition of this polysemic term is not easy which makes it difficult to regulate. In 2018, the Code of Practice on Disinformation was adopted. The Code is a soft law tool described as a voluntary, self-regulatory mechanism composed of various content moderation commitments such as developing clear policies regarding the identity and misuse of automated bots and closing false accounts; investing in technologies to help internet users to 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. In September 2020, the European Commission published its assessment of the Code of Practice on Disinformation. Numerous positive impacts have been found but also a serious number of shortcomings. The lack of key definitions, vague concepts, a narrow scope, combined with lack of enforcement and monitoring mechanisms undermined the Code's impact and its potential for being a level playing field instrument[19]. 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 in 2022, with 34 signatories who have joined the revision process of the 2018 Code. Until now the EU regulation efforts were quite cautious with a self-regulation approach but the EU is stepping up its effort with the revised version closely tied with the DSA[20]. The question of the Code's commitments relationship with the DSA is still to be clarified in practice similarly to some concepts such as harmful disinformation[21]. The combination of the Code and the DSA brings new obligations for VLOPS such as the systemic risks assessment and mitigation measures which could also cover disinformation. The body responsible for monitoring the compliance with the Code is the EC. Some have criticised whether it has enough staff and resources to conduct this mission and whether a political institution is the best suited to decide on disinformation.

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

3

Alternative Approaches and Future Trends in Content Moderation

Content moderation is a multifaceted coin composed of divergent approaches by private actors and platforms. Given the massive 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 future cascade content moderation initiatives or mechanisms.

End-user Moderation or Community-led Moderation

In the following, we discuss initiatives for end-user or community-led moderation. For more information, we refer the reader to D6.2, Section 4.1, pp. 79-89.

Self-moderated communities

- **Wikipedia:** Wikipedia is a volunteer community moderated platform with over 300 language versions. Content moderation on Wikipedia is carried out by volunteers consisting of administrators, editors, bots, and monitoring tools. The rules of content moderation on Wikipedia may differ in different language versions. There are no moderators or automated content recognition tools governed by the platform itself, but the volunteer community is responsible for moderation. Contributors are legally responsible for all contributions and edits, and prohibited from uploading content which includes defamation, harassment, threatening, and copyright-infringing content. However, being a self-moderated community platform does not rule out issues arising from content moderation, and content on Wikipedia could still be biased and inaccurate. The relation between Wikipedia and copyright protected content on the one hand and with the DSA on the other hand are also investigated in this section.
- **Discord:** Discord is a community content moderation platform that relies on server admins to handle moderation. The platform offers a recently introduced moderation tool called 'Auto Mod' to assist admins

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

and moderators in keeping their servers safe. Nevertheless, Discord faced challenges in the past with extremist users and groups utilising the platform to spread harmful content and organise attacks, resulting in some concerns about the effectiveness of Discord's community moderation model.

End-users have generally higher confidence in distributed moderation than centralised moderation as the moderators are closer to them[22]. But a number of challenges arise such as the lack of relevant expertise, the personal biases, lack of consistency, potential discrimination of minorities, the burden of such a task on community moderators[23]. When it comes to the DSA, from the first look it doesn't seem that community content moderation falls in the scope of the DSA but it doesn't waive Wikipedia's and Discord's obligations based on the legislation[24].

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[25]. The servers - generally called "instances" - are 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 and therefore choose what content will flow and what content will be blocked. No central authority can decide which instances are valid but users have the ability to switch instances if dissatisfied and move their account data with them. Therefore, 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.

→ **Mastodon** is the largest federated social network. Each Mastodon instance chooses its own content moderation policies creating a whole variety of them with some more or less restrictive. Mastodon will likely be categorised under the DSA as a 'hosting service' and will need to comply with the relevant set of rules[26].

Content Moderation in 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"[27]. Facebook's VR Metaverse is just one example of such a metaverse world. With great opportunities in the metaverse come great risks. They raise 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. Some of these risks have already materialised as researchers found 100 potential violations of Facebook's policies for VR in 11 hours and 30 minutes of recordings of user behaviour in the app[28]. 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 which have serious technical limitations already explained above. The most serious risk is perhaps the lack of understanding of the context. Slight behavioural changes or the use of symbols that exploit the algorithms' lack of comprehension of context may go undetected[29]. Other platforms may choose to adopt more of a decentralised approach that allows communities and volunteers to moderate the content. However, community-led moderation can lead to a lack of platform-wide standards and human moderators' burnouts[30]. When it comes to the DSA, the topic of virtual reality is not specifically addressed in the text, and it will be necessary to define clearly in which dimension does this fall in the scope. The question whether some AI Act proposal provisions could be applicable to metaverse remains open and more clarity should be brought on that aspect in the negotiations.

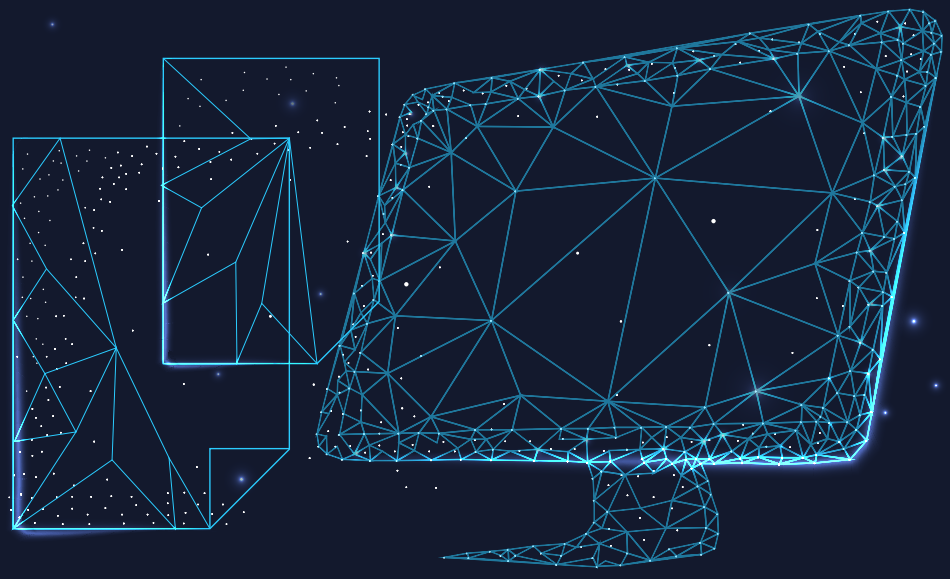
Accountability Initiatives

Online platforms or civil society started to initiate alternative accountability initiatives. Self-regulation initiatives seem to flourish in the content moderation landscape and provide some interesting concepts to study. 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 policymaking asking more from platforms to protect fundamental rights. For more information, we refer the reader to D6.2, Section 4.2, pp. 89-98.

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

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 can be considered as a self-regulation mechanism. 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. The Board determines if Meta's decisions were made in accordance with Meta's stated values and policies. Only a few cases are actually taken and reviewed by the board[31]. In addition, the OB can issue non-binding recommendations about the platform's policies. The OB is a controversial institution and has both supporters and critics. On the one hand the OB has contributed to improved transparency of Meta Content moderation decisions, overruled some of the most problematic Meta's decisions and issued far-reaching recommendations. On the other hand, it lacks diversity in its staff and has only a limited impact based on its mandate even if things are about to change[32]. The OB is presenting similarities with some newly adopted DSA obligations without matching their scope[33]. The OB will need to undergo structural and drastic changes to be able to fit as an internal complaint system or an out of court dispute settlement, which constitute unlikely scenarios. The OB can still, if improved, become a valuable complement to robust, international legislation. It has set a trend for other platforms leading to the creation of the Twitter Trust and Safety Council, the TikTok Content Advisory Council, the Spotify Safety Advisory Council, and Twitch's Safety Advisory Council.



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

Social Media Councils

In 2018, Article 19 (civil society), suggested exploring a new model of effective self-regulation for social media: social media councils (SMC). 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[34]. SMC would become a multistakeholder, transparent, inclusive accountability mechanism for content moderation on social media. They could fulfil the following 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[35]. There are both opportunities and challenges which come which such a model (Figure 2).

Challenges with current practices of content moderation	Advantages of an SMC
Antagonism between stakeholders	Acts as a forum for cooperation and co-learning
No esternal oversight of content moderation decisions	External oversight based on international human rights law
No remedy for individual users	Individual users have acces 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 2 - 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)

A. Kuczerawy gives a closer look at the SMC and the DSA[36]. She investigates whether these models could be considered as an internal complaint mechanism or an out-of-court dispute settlement. Even if the SMC would not constitute an internal complaint mechanism, they could be a fit as an out-off-court dispute settlement.

Section 4 concludes that "neither pure self-regulation nor aggressive government regulation seems likely to cover all the challenges digital platforms face"[37]. For self-regulation to be effective, it cannot happen exclusively at the platforms' level. Some widely accepted set of rules or codes of conduct are welcome. However, regulatory pressure appears necessary[38]. To solve content moderation challenges, private and public actors will have to collaborate, which is a trend already visible with the EU revised Code of practice on disinformation and the DSA.

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

AI4Media Workshop on AI and Content Moderation

On February 6, 2023, KUL and UvA organised a workshop to explore the challenges faced by the industry on AI and content moderation. The workshop discussed the main challenges faced by those either building AI systems for content moderation or using these systems. The participants of the workshop were diverse, and included companies developing image recognition solutions, consultancies doing content moderation analysis, an AI4Media-funded project focusing on robust and adaptable comment filtering, newspapers, a major platform, and technology companies. The workshop was held under the Chatham House Rule and was an invitation-only event. The workshop was held under the Chatham House Rule and was an invitation-only event.

Prior to the meeting, participants were asked to fill in a short survey where they were asked to identify the top three challenges they are currently battling with in their daily work on AI in content moderation.

The main challenges identified in the workshop were the following:

- Lack of access to training data,
- Lack of transparency in AI models,
- Ensuring human oversight in real-time moderation,
- Defining and classifying hate speech and toxicity in a context-sensitive way, and
- Lack of inclusivity, such as minor languages not properly being represented.

The main takeaways from the workshop were:

- AI is a tool for content moderation and should not fully replace human review.
- Impact of content moderation on human reviewers should be taken into account.
- More attention should be given to fine-tuning AI models and reducing noise.
- Evaluation methods, processes and criteria for models should be established to evaluate the positive and negative impacts of the models.

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

- Dislocation of content moderation is a concern.
- The work of human rights workers, archivists, and historians should be considered in content moderation.
- Participants have raised the question of whether content that is removed could be considered public domain information and if a right to request already moderated data should be established.
- Respect for the GDPR is often used as an excuse not to share data on removed content.
- Content moderation should be open to a variety of players at different levels of the chain.
- Small and midsize companies face challenges in content moderation.
- An advisory board on content moderation and AI would be helpful for stakeholders' interest.
- Some content moderation subjects are overlooked, including fraud, direct incitement to violence, self-harm, and crime plotting.

For more information about the workshop and its outcomes, we refer the reader to D6.2, Section 5, pp. 99-102.

4.

Policy Recommendations on Content Moderation

The deliverable 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 (See D6.2, Section 6, pp. 103-110).

Horizontal and High-level Recommendations

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 line with the gap analysis and the identified needs, we propose the below high-level policy recommendations for content moderation:

- 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 regular updates of the terms of use, and 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).
- Ensure proper processes of data access for research, historical, archival, and 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 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.

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

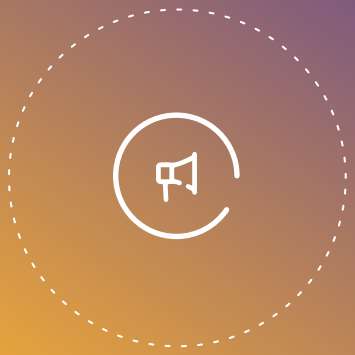
Problem-specific Recommendations

A set of problem-specific recommendations was also developed focusing on specific types of content such as terrorist content, copyright-protected content, child sexual abuse material, hate speech, and disinformation.

Type of contents	Recomendations
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. 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.

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

5
Conclusion



Content moderation being 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. 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. With new technological advances, come new benefits, but also potential new risks for fundamental rights. As shown 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 (legislative or non-legislative) 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.

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

6

References

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

2. '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.

3. 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.

4. 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.

5. 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).

6. 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; Robert Gorwa, 'What Is Platform Governance?' (2019) 22 Information, Communication & Society 854

7. 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'. Available here. Accessed 16 November 2022.

8. 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

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

10. 'Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis' (Center for Democracy and Technology, 20 May 2021). Available here. Accessed 31 January 2023.

11. 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.

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

12. 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

13. 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).

14. 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.

15. Figure 1 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). Available here. Accessed 23 January 2023.

16. 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

17. Charlotte Somers, 'The Proposed CSAM Regulation: Trampling Privacy in the Fight against Child Sexual Abuse?' (CITIP blog, 3 January 2023) accessed 20 January 2023.

18. Barbora Bukovská, 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online'; EPRS, Polarisation and the use of technology in political campaigns and communication.

19. 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) accessed 20 March 2023.

20. 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) accessed 20 March 2023.

21. Natali Helberger and others, 'The EU's regulatory push against disinformation: What happens if platforms refuse to cooperate?,' (VerfBlog, 2022/8/05), accessed February 20, 2023. ; 'Position of the EU DisinfoLab on the 2022 Code of Practice on Disinformation' (EU DisinfoLab, September 8, 2022), accessed February 20, 2023

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

23. 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 Ask Historians' (2020) 4 Proceedings of the ACM on Human-Computer Interaction 19:1.; Stefanie Duguay and others, 'Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine' (2020) 26 Convergence 237; 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 HumanComputer Interaction 32:1

24. Dimi Dimitrov, 'DSA: Political Deal Done' (Free Knowledge Advocacy Group EUApril 26, 2022) accessed February 25, 2023.

25. Alan Z. Rozenshtein 'Moderating the Fediverse: Content Moderation on Distributed Social Media', SSRN accessed 24 January 2023.

26. Konstantinos Komaitis, 'Can Mastodon Survive Europe's Digital Services Act?' (Tech Policy Press, 16 November 2022) accessed 26 January 2023.

27. EPRS, Metaverse. Opportunities, risks and policy implications

28. Eli Cohen Lawson, 'New Research Shows Metaverse Is Not Safe for Kids' (Center for Countering Digital Hate | CCDH, 30 December 2021) accessed 26 January 2023

29. EPRS, Metaverse. Opportunities, risks and policy implications

30. Juan Londoño 'Lessons from Social Media for Creating a Safe Metaverse' accessed 26 January 2023

31. 'Oversight Board Cases' (Meta Transparency Centre) accessed 20 March 2023

32. 'Oversight Board Announces Plans to Review More Cases, and Appoints a New Board Member' accessed 20 March 2023

33. David Wong and Luciano Floridi, 'Meta's Oversight Board: A Review and Critical Assessment' [2022] Minds and Machines accessed 21 December 2022

34. Stefanie Barth, 'Can Social Media Councils Tame Digital Platforms? – Digital Society Blog' (HIIG, 29 September 2022) accessed 1 March 2023

35. Article 19, 'Social Media Councils, One Piece in the Puzzle of Content Moderation' (2021)

36. 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 here.

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

38. 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