Online Platforms' Moderation of Illegal Content Online

DE STREEL, Alexandre; Defreyne, Elise; Jacquemin, Herve; Ledger, Michele; Michel, Alejandra

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

STUDY

Requested by the IMCO committee

European Parliament

Online Platforms' Moderation of Illegal Content Online

Law, Practices and Options for Reform

Policy Department for Economic, Scientific and Quality of Life Policies
Directorate-General for Internal Policies
Authors: Alexandre DE STREEL et al.
PE 652.718 - June 2020

EN

Online Platforms' Moderation of Illegal Content Online

Law, Practices and Options for Reform

Abstract

Online platforms have created content moderation systems, particularly in relation to tackling illegal content online. This study reviews and assesses the EU regulatory framework on content moderation and the practices by key online platforms. On that basis, it makes recommendations to improve the EU legal framework within the context of the forthcoming Digital Services Act.

This document was provided by the Policy Department for Economic, Scientific and Quality of Life Policies at the request of the committee on Internal Market and Consumer Protection (IMCO).

This document was requested by the European Parliament's committee on Internal Market and Consumer Protection.

AUTHORS
UNIVERSITY OF NAMUR (CRIDS/NADI): Alexandre DE STREEL, Elise DEFREYNE, Hervé JACQUEMIN, Michèle LEDGER, Alejandra MICHEL
VVA: Alessandra INNESTI, Marion GOUBET, Dawid USTOWSKI

ADMINISTRATOR RESPONSIBLE
Christina RATCLIFF

EDITORIAL ASSISTANT
Roberto BIANCHINI

LINGUISTIC VERSIONS
Original: EN

ABOUT THE EDITOR
Policy departments provide in-house and external expertise to support EP committees and other parliamentary bodies in shaping legislation and exercising democratic scrutiny over EU internal policies.

To contact the Policy Department or to subscribe for email alert updates, please write to:
Policy Department for Economic, Scientific and Quality of Life Policies
European Parliament
L-2929 - Luxembourg
Email: Poldep-Economy-Science@ep.europa.eu

Manuscript completed: June 2020
Date of publication: June 2020

This document is available on the internet at:
http://www.europarl.europa.eu/supporting-analyses

For citation purposes, the study should be referenced as: De Streel, A. et al., *Online Platforms' Moderation of Illegal Content Online*, Study for the committee on Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2020.

CONTENTS

LIST OF ABBREVIATIONS

ADR	Alternative Dispute Resolution
AI	Artificial Intelligence
AVMSD	Audio-Visual Media Service Directive
BEUC	European Consumer Organisation
CCIA	Computer and Communications Industry Association
CDA	Communications Decency Act
CDSMD	Copyright in the Digital Single Market Directive
CDT	Center for Democracy & Technology
CEN	Comité Européen de Normalisation – European Committee for Standardisation
CENELEC	Comité Européen de Normalisation en Electronique et en Electrotechnique – European Committee for Electrotechnical Standardisation
CEO	Chief Executive Officer
CEP	Counter Extremism Project
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
CPC	Consumer Protection Cooperation
CRFD	Counter-Racism Framework Decision
CSO	Civil Society Organisation
CSAED	Child Sexual Abuse and Exploitation Directive
CSAM	Child Sexual Abuse Material
CTD	Counter-Terrorism Directive
DSA	Digital Services Act
ECD	e-Commerce Directive
ECHR	European Court of Human Rights
EDiMA	European Digital Media Association
ERGA	European Regulators Group for Audio-Visual Media Services
ETSI	European Telecommunications Standards Institute
EU	European Union
EuroISPA	European Internet Services Providers Associations
FRA	European Union Agency for Fundamental Rights
GDPR	General Data Protection Regulation
HLEG	High-Level Expert Group

HSP	Hosting Service Provider
ICT	Information and Communication Technologies
IPR	Intellectual Property Rights
ISP	Internet Service Provider
KPI	Key Performance Indicator
MoU	Memorandum of Understanding
MS	Member State
N&A	Notice-and-Action
NCMEC	National Center for Missing and Exploited Children (US)
NetzDG	German Network Enforcement Act
NGO	Non-Governmental Organisation
OFCOM	The regulator and competition authority for the UK communications industries
PSCSP	Public Space Content-Sharing Platform
SMC	Social Media Council
TERREG	Proposal for a Regulation on preventing the dissemination of terrorist content online
UK	United Kingdom
UN	United Nations
URL	Uniform Resource Locator
US	United States
VSP	Video-Sharing Platforms
WHO	World Health Organization

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

EU regulatory framework on online content moderation

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 and the type of content reflecting a risk-based approach. The e-Commerce Directive of 2000 contains the baseline regime applicable to all categories of platforms and all types of content. The Directive provides the following rules: (i) the 'country of origin' principle, which is the cornerstone of the Digital Single Market; (ii) an exemption of liability for hosting platforms which remain passive and neutral and which remove the illegal content online as soon as they are made aware of it; (iii) the prohibition of general monitoring measures to protect fundamental rights; and (iv) the promotion of self- and co-regulation as well as alternative dispute resolution mechanisms.

This baseline regulatory regime has been complemented in 2018 by the revised Audio-Visual Media Services Directive, which imposes more obligations to one category of online platforms, the **Video-Sharing Platforms**. **They should take appropriate and proportionate measures**, preferably through co-regulation, in order to protect the general public from illegal content (terrorist content, child sexual abuse material, racism and xenophobia or other hate speech), and to protect minors from harmful content. Those measures must be **appropriate** in the light of the nature of the content, the category of persons to be protected and the rights and legitimate interests at stake and be **proportionate** taking into account the size of the platforms and the nature of the provided service.

Those rules are then **strengthened by stricter rules for four types of content for which illegality has been harmonised at the EU level**:

- first, the Counter-Terrorism Directive defines the public provocation to commit a terrorist offence and requires, following transparent procedures and with adequate safeguards, **Member States to take removing and blocking measures against websites containing or disseminating terrorist content**. The European Commission aims to go further and has made a proposal, which has not yet been adopted by the EU co-legislators, for a regulation which would require **hosting services providers to take measures to remove terrorist content**;
- second, the **Child Sexual Abuse and Exploitation Directive** defines child pornography and requires, following transparent procedures and with adequate safeguards, **Member States to take removing and blocking measures against websites containing or disseminating child sexual abuse material**;
- third, the Counter-Racism Framework Decision provides that **Member States must ensure that racist and xenophobic hate speech is punishable**, but does not impose detailed obligations related to online content moderation practices;
- fourth, the Copyright in Digital Single Market Directive establishes a **new liability regime for online content-sharing platforms**; they must conclude an agreement with the rights-holders for the exploitation of the works and, if they fail to do so, they are liable for the content violating copyright on their platforms unless they make their best effort to alleviate such violations.

Those stricter rules imposed by EU hard-law are all **complemented by self-regulatory initiatives** agreed by the main online platforms, often at the initiative of the European Commission. They contain a range of commitments, some of which are directly related to content moderation practices and others which support such practices. However, the evaluation of those initiatives shows **difficulties in measuring the commitments taken** and in reporting on their effectiveness.

With regard **online disinformation**, which is not always illegal but can be very harmful to EU values, the main platforms have agreed to a Code of Practice in 2018. Such Code is closely monitored by the European Commission.

In addition to this multi-layered EU regulatory framework, **several Member States have adopted national rules on online content moderation**, in particular for hate speech and online disinformation. The legal compatibility of those national initiatives with the EU legal framework is not always clear and the multiplication of national laws **seriously risks undermining the Digital Single Market**.

Online Content Moderation Practices and their effectiveness

Online platforms, big and small, **rely on Terms of Service/Terms of Use or Community Standards/Guidelines to regulate and user behaviour and base their illegal content online moderation practices**. These Terms and Standards/Guidelines do not necessarily reflect a specific legal system. However, as they are designed to prevent harm, online platforms' policies do overlap in several instances with local law. These private Codes of Conduct implemented by online platforms may vary from one country to another; they are often **stricter in identifying illegal content online to be removed than national laws** or jurisdictions within which they provide their services.

The **main tools used by online platforms to identify illegal content online are 'notice-and-takedown'/flagging by users, keywords/filters and AI tools based machine learning models**. Most platforms noted that depending on the type of illegal content online, automated tools have their limits in terms of accuracy, and thus, frequently must be accompanied by pre-/post-human moderation to ensure accuracy. The majority of online platforms have argued that the policies put in place by them to moderate illegal content online contribute to **reducing the aggressive nature and the quantity of illegal content online**.

All online platforms interviewed have implemented **transparency policies** on how they operate and respect fundamental rights. Moreover, all of them have **complaint mechanisms** in place for their users to report on illegal content online. However, some platforms have emphasised that many user complaints are off-topic or unsubstantiated and consequently unactionable. Almost all online platforms interviewed allow users to **appeal** against their decisions on the moderation of illegal content online through a 'counter-notice' procedure.

However, most of the interviewed NGOs, trade/industry associations and hotlines reporting illegal content online stated that the measures used by online platforms **are not sufficiently effective in moderating illegal content online and in striking an appropriate balance with fundamental human rights**. Most NGOs and hotlines reporting illegal content online have argued that the effectiveness of the measures deployed by platforms to enable users to report illegal content fluctuates according to the online platform. Additionally, they noted that access to 'notice-and-takedown' procedures is not always user-friendly, whereas they should be easily accessible and not hidden in obscurity.

The **main challenges in moderating illegal content online are linked to the large quantity of online content** on platforms, which makes it difficult for users, regulators or moderators to assess all **content as well as the fragmentation of laws regarding online content**. The Member States are free to set their own rules regarding illegal content online, which limits the efficiency of platforms that have to create country-specific processes accordingly. The lack of a common definition of "illegal content" also makes the moderation by platforms more complex as Member States may refer to different definitions. Therefore, some online platforms mentioned that this places the burden on them to

identify the intent of the content uploader, which might incentivise online platforms to block lawful content in case of doubt on the illegality of the content.

Several stakeholders also note that the **current legislative framework on content moderation focuses mainly on the responsibility of online platforms**, while they argue that this should be balanced with rights and obligations of other stakeholders. Most NGOs and industry/trade associations interviewed disagreed with the idea of specific duty of care regimes. They pointed out that **new statutory obligations to remove illegal content online should apply horizontally to any type of illegal content** to avoid regulatory fragmentation.

Almost all online platforms interviewed considered the terms of the **existing liability principles of intermediary service providers** of the e-Commerce Directive as **fit-for-purpose**. However, two indicated that the concept of 'active' and 'passive' intermediary service providers should be reformed. This is because the concept may no longer adequately reflect the economic, social, and technical reality of current services across their lifecycle. Most platforms mentioned that the limitation of liability for Internet intermediaries is a good solution. This is because it allows for protection of fundamental rights, the rule of law and the open Internet.

Regarding **the solutions** to improve the moderation of illegal content online by platforms, stakeholders suggested to put in place **harmonised and transparent 'notice-and-action' processes**. Some stakeholders suggested **strengthening the networks of fact-checkers and hotlines** across the EU. **In terms of fundamental rights, several stakeholders recommended to enforce existing EU rules** and to make them more consistently interpreted across Member States.

Almost all NGOs, industry/trade associations and online platforms interviewed consider that the **existence of different content moderation practices in EU Member States hinder the fight against illegal content online**. Several online platforms stressed that a harmonised approach would enable service providers to have more clarity on what they must do to fulfil their legal responsibilities, while upholding fundamental rights.

In the context of **safeguarding fundamental rights**, most NGOs noted that online platforms' moderating practices should increase moderation transparency, access to data, and information regarding platforms' decision-making processes. In addition, they should ensure human review of the decisions on the user-generated content and contextual expertise. Some online platforms have acknowledged that platforms' incentive to over-remove legal content constitutes the most considerable **threat of unjustified interference** with fundamental rights.

International benchmarking

The analysis conducted in six countries/regions of the world shows that **all of these countries have a regulatory and policy framework related to illegal content online**. The policies may relate either to online hate speech, defamation, child sexual abuse material, copyright infringements or online disinformation. The large majority of the countries investigated have regulatory measures in place related to at least one or several of these areas. Some **policy guidelines have also been established** by some countries to guide online platforms when moderating illegal content online. In almost all countries covered, NGOs and academics have recommended ways to improve the online content moderation practices.

In addition to the regulatory and policy measures put in place to frame illegal content online, **most of the online platforms worldwide apply their own Terms of Service/Terms of Use to moderate online content**, which often stem from the large US online platforms.

In comparison to the set of measures identified in the various countries, the **EU seems to have one of the most developed regulatory frameworks related to illegal content online and its moderation by online platforms**.

Policy Recommendations for the Digital Services Act

The revised EU regulatory framework for online content moderation, which will result from the **forthcoming Digital Services Act, could be based on the following objectives and principles**:

- sufficient and effective safeguards to protect fundamental rights;
- a strengthening of the Digital Single Market;
- a level playing field between offline and online activities;
- technological neutrality;
- incentives for all stakeholders to minimise the risk of errors of over and under removal of content;
- proportionality of the potential negative impact of the content and the size of the platforms; and
- coherence with existing content-specific EU legislation.

The **baseline regulatory regime** applicable to all types of content and all categories of platforms could strengthen in an appropriate and proportionate manner the responsibility of the online platforms to ensure a safer Internet. To do that, it could include a **set of fully harmonised rules on procedural accountability** to allow public oversight of the way in which platforms moderate content. Those rules could include: (i) common EU principles to improve and harmonise the **'notice-and-takedown' procedure** to facilitate reporting by users; (ii) the encouragement for the platforms to take, where appropriate, proportionate, specific **proactive measures** including with automated means; and (iii) the strengthening of the **cooperation with public enforcement authorities**. Those new rules could be based on the measures recommended by the European Commission in its 2018 Recommendation on measures to effectively tackle illegal online content as well as on the measures imposed on Video-Sharing Platforms by the revised 2018 Audio-Visual Media Services Directive.

This baseline regulatory regime could be complemented with **stricter rules imposing more obligations, when the risk of online harm is higher**. Stricter rules are already imposed **according to the type of content**: more obligations are imposed for the moderation of the online content with the highest potential negative impact on the society such as terrorist content, child sexual abuse material, racist and xenophobic hate speech and some copyright violations. Stricter rules could also be imposed **according to the size of the platform**: more obligations could be imposed on the platforms whose number of users is above a certain threshold, which could be designated as Public Space Content-Sharing Platforms (PSCSPs).

As often in EU law, enforcement is the weak spot and therefore, the forthcoming Digital Services Act should ensure that any online content moderation rule is **enforced effectively**. Such enforcement should be ensured **by public authorities**, in particular regulatory authorities and judicial courts. The 'country of origin' principle should be maintained, hence the **online platforms should in principle be supervised by the authorities of the country where they are established**. However, the authorities of the country of establishment may not have sufficient means and incentives to supervise the largest platforms; hence, an **EU authority could be set up to supervise the PSCSPs**. In addition, the enforcement could be improved with, on the one hand, a **better coordination between national**

authorities by relying on the Consumer Protection Cooperation Network and, on the other hand, **better information disclosure** in the context of Court proceedings.

Given the massive explosion of online content, **public authorities may not be sufficiently well-geared to ensure the enforcement of content moderation rules and may need to be complemented with private bodies**. Those could be the platforms themselves, self-regulatory bodies or co-regulatory bodies. The involvement of private bodies seems inevitable, but should not lead to full delegation of State sovereign power to private firms or a privatisation of the public interest, hence co-regulation could be an effective tool, preferred to self-regulation.

Next to specific obligations regarding the moderation of illegal content online, **complementary broader measures are also necessary** such as more transparency on the way moderation is done and support to journalists, Civil Society Organisation or NGOs, which contribute to the fight against illegal content.

1. SCOPE AND OBJECTIVES OF THIS STUDY[1]

Background

Online platforms have created content moderation systems, particularly in relation to tackling illegal content online. Such moderation practices can include depublication, delisting, downranking and can lead to some forms of censorship of information and/or user accounts from social media and other online platforms. Those practices may have an impact on the overall quality and quantity of online content. They are usually based on an alleged violation of online platforms' Community Standards/Guidelines policies. Whether faced with online disinformation, harassment, or violence, content moderators have a fundamental role to play in online platforms. However, the tools that online platforms use to curb trolling, ban hate speech, or restrict pornography can also silence content relevant to the public. In recent years, content moderation practices have become a matter of intense public interest. Also, the co-existence of different moderation practices at national level could affect the functioning of the Internal Market.

In its 2020 *Digital Strategy Communication*, the European Commission noted that "*it is essential that the rules applicable to digital services across the EU are strengthened and modernised, clarifying the roles and responsibilities of online platforms. The sale of illicit, dangerous or counterfeit goods, and dissemination of illegal content must be tackled as effectively online as it is offline*". Therefore, the Commission announced as one key action part of the Digital Services Act (DSA) package: "*new and revised rules to deepen the Internal Market for Digital Services, by increasing and harmonising the responsibilities of online platforms and information service providers and reinforce the oversight over platforms' content policies in the EU*"[2].

Aim of the Study

This study analyses the current EU legal framework on online content moderation, the practices of key online platforms active in the EU and, on that basis, makes recommendations for reforms.

The study is organised as follows: after this introduction, Section 2 reviews the EU legal framework for online content moderation and also briefly summarises interesting national laws or initiatives (in Germany, France and the UK). Section 3 provides an overview of the online content moderation practices on the basis of interviews of online platforms and other stakeholders. Section 4 reviews briefly some policies and recommendations on online content moderation practices made in other regions of the world. On the basis of all the information included in the previous sections, Section 5 makes recommendations to improve the EU legal framework on online content moderation.

[1] The authors want to thank Aleksandra Kuczerawy for her very helpful comments and suggestions.

[2] Communication from the Commission of 19 February 2020, Shaping Europe's digital future, COM(2020) 67, pp.11-12. See also the European Commission Evaluation Roadmap/Inception Impact Assessment of 3 June 2020 on the Digital Services Act package: deepening the Internal Market and clarifying responsibilities for digital service, available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-Internal-Market-and-clarifying-responsibilities-for-digital-services.

2. EU REGULATORY FRAMEWORK ON ONLINE CONTENT MODERATION

KEY FINDINGS

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 and the type of content reflecting a risk-based approach. The e-Commerce Directive of 2000 contains the baseline regime applicable to all categories of platforms and all types of content. The Directive provides the following rules: (i) the 'country of origin' principle, which is the cornerstone of the Digital Single Market; (ii) an exemption of liability for hosting platforms which remain passive and neutral and which remove the illegal content online as soon as they are made aware of it; (iii) the prohibition of general monitoring measures to protect fundamental rights; and (iv) the promotion of self- and co-regulation as well as alternative dispute resolution mechanisms.

This baseline regulatory regime has been complemented in 2018 by the revised Audio-Visual Media Services Directive, which imposes more obligations to one category of online platforms, the **Video-Sharing Platforms**. **They should take appropriate and proportionate measures**, preferably through co-regulation, in order to protect the general public from illegal content (terrorist content, child sexual abuse material, racism and xenophobia or other hate speech), and to protect minors from harmful content. Those measures must be **appropriate** in the light of the nature of the content, the category of persons to be protected and the rights and legitimate interests at stake and be **proportionate** taking into account the size of the platforms and the nature of the provided service.

Those rules are then **strengthened by stricter rules for four types of content for which illegality has been harmonised at the EU level**:

- First, the Counter-Terrorism Directive defines the public provocation to commit a terrorist offence and requires, following transparent procedures and with adequate safeguards, **Member States to take removing and blocking measures against websites containing or disseminating terrorist content**. The European Commission aims to go further and has made a proposal, which has not yet been adopted by the EU co-legislators, for a regulation which would require **hosting services providers to take measures to remove terrorist content**;
- Second, the **Child Sexual Abuse and Exploitation Directive** defines child pornography and requires, following transparent procedures and with adequate safeguards, **Member States to take removing and blocking measures against websites containing or disseminating child sexual abuse material**;
- Third, the Counter-Racism Framework Decision provides that **Member States must ensure that racist and xenophobic hate speech is punishable**, but does not impose detailed obligations related to online content moderation practices;
- Fourth, the Copyright in Digital Single Market Directive establishes a **new liability regime for online content-sharing platforms**; they must conclude an agreement with the rights-holders for the exploitation of the works and, if they fail to do so, they are liable for the content violating copyright on their platforms unless they make their best effort to alleviate such violations.

Those stricter rules imposed by EU hard-law are all **complemented by self-regulatory initiatives** agreed by the main online platforms, often at the initiative of the European Commission. They contain a range of commitments, some of which are directly related to content moderation practices and others which support such practices. However, the evaluation of those initiatives shows **difficulties in measuring the commitments taken** and in reporting on their effectiveness.

With regard **online disinformation**, which is not always illegal but can be very harmful to EU values, the main platforms have agreed to a Code of Practice in 2018. Such Code is closely monitored by the European Commission.

In addition to this multi-layered EU regulatory framework, **several Member States have adopted national rules on online content moderation**, in particular for hate speech and online disinformation. The legal compatibility of those national initiatives with the EU legal framework is not always clear and the multiplication of national laws **seriously risks undermining the Digital Single Market**.

2.1. Definition of illegal content online

EU law makes illegal four types of content: (i) child sexual abuse material; (ii) racist and xenophobic hate speech; (iii) terrorist content; and (iv) content infringing Intellectual Property Rights (IPR). Beyond those four types, there is no EU harmonisation of the illegal content online. Thus, the same type of content may be considered illegal, legal but harmful or legal and not harmful across the Member States. This study distinguishes between online content that is illegal under EU law and a residual category, which includes content that may be illegal under national law.

2.1.1. Online content illegal under EU law

The **Counter-Terrorism Directive (CTD)**[3] defines **the public provocation to commit a terrorist offence**[4] in broad terms as it covers indirect advocacy and ill-defined 'glorification' of terrorist acts. It only requires that the conduct causes a danger that the offences may be committed. It does not require that the conduct creates an actual risk or an imminent danger of harm, which would be harder to establish. In 2018, the European Commission **proposed a Regulation on preventing the dissemination of terrorist content online (TERREG) to complement the CTD**, in particular by tackling the misuse of hosting services for terrorist purposes[5]. The Commission proposes to define terrorist content as covering material that incites or advocates the commission of terrorist offences, encourages the contribution to terrorist offences, promotes the activities of a terrorist group or provides methods and techniques for committing terrorist offences[6]. However, such definition has

[3] Directive 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism, OJ [2017] L 88/6.

[4] CTD, Article 5 defines the public provocation to commit a terrorist offence as "the distribution, or otherwise making available by any means, whether online or offline, of a message to the public, with the intent to incite the commission of [terrorist offences], where such conduct, directly or indirectly, such as by the glorification of terrorist acts, advocates the commission of terrorist offences, thereby causing a danger that one or more such offences may be committed".

[5] Proposal of the European Commission of 12 September 2018 for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online, COM (2018) 640.

[6] TERREG Proposal, Article 2(5) defines terrorist content as: "(a) inciting or advocating, including by glorifying, the commission of terrorist offences, thereby causing a danger that such acts be committed; (b) encouraging the contribution to terrorist offences; (c) promoting the activities of a terrorist group, in particular by encouraging the participation in or support to a terrorist group within the meaning of CTD; (d) instructing on methods or techniques for the purpose of committing terrorist offences".

been criticised for being too broad and including legitimate forms of expression (e.g. journalists and NGO reports on terrorist content)[7] and the European Parliament suggests to align the definition of terrorist content in TERREG with the definition contained in the CTD[8].

The Child Sexual Abuse and Exploitation Directive (CSAED) establishes minimum rules concerning the definition of criminal offences and sanctions in the area of child sexual exploitation and abuse[9]. The Directive **provides for a broad definition of Child Sexual Abuse Material (CSAM)** that includes real child pornography that visually depicts a child engaged in real or simulated sexually explicit conduct or virtual child pornography, i.e. computer-generated pornographic material involving children[10].

The **Counter-Racism Framework Decision (CRFD)**, which was adopted by the Council alone, seeks to combat particularly serious forms of racism and xenophobia through criminal law but does not define racism and xenophobia nor use the terms **racist and xenophobic hate speech**[11]. Instead, the CRFD criminalises two types of speech - publicly inciting to violence or hatred and publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes - when they are directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin. The list of protected grounds is limited to these five characteristics.

With regard to **Intellectual Property Rights**, some harmonisation has been achieved at the EU level through different EU Directives such as the InfoSoc Directive[12], which harmonises the rights of reproduction, distribution and communication to the public and the legal protection of anti-copying devices and rights management systems.

2.1.2. Illegal content online under national law

While only the hate speech which is racial and xenophobic has been made illegal by the CRFD, Member States may go beyond the EU minimum and criminalise **other types of hate speech**, by referring to a broader list of protected characteristics (a list including e.g. religion, disability, sexual orientation). Article 21 of the EU Charter of Fundamental Rights can serve as a reference point for the Member States in that it prohibits all forms of discrimination with regard to a detailed list of protected characteristics[13]. Also, Member States can be guided by a recent Resolution of the European Parliament, which condemns hate crime and speech by bias against a person's disability, sexual orientation, gender

---

7 Opinion 2/2019 of the European Union Agency for Fundamental Rights 12 February 2019 on the Proposal for a Regulation on preventing the dissemination of terrorist content online and its fundamental rights implications.

8 LIBE Committee Report of April 2019 on the proposal for a regulation on preventing the dissemination of terrorist content online (C8-0405/2018), AM 52-57.

9 Directive 2011/92 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, O.J. [2011] L 335/1. This study uses the terms Child Sexual Abuse Material, rather than child pornography, following the Resolution of the European Parliament of 11 March 2015 on child sexual abuse online, point 12 and the Resolution of the European Parliament of 14 December 2017 on the implementation of Directive 2011/93 on combating the sexual abuse and sexual exploitation of children and child pornography, point 4.

10 CSAED, Article 2: child pornography means: "(i) any material that visually depicts a child engaged in real or simulated sexually explicit conduct; (ii) any depiction of the sexual organs of a child for primarily sexual purposes; (iii) any material that visually depicts any person appearing to be a child engaged in real or simulated sexually explicit conduct or any depiction of the sexual organs of any person appearing to be a child, for primarily sexual purposes; or (iv) realistic images of a child engaged in sexually explicit conduct or realistic images of the sexual organs of a child, for primarily sexual purposes." See also Jeney (2015).

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

12 Directive 2001/29 of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, O.J. [2001] L167/10.

13 EU Charter of Fundamental Rights, Article 21(1): "Any discrimination based on any ground such as 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 shall be prohibited".

identity, sex characteristics or minority status. Finally, Member states must also take into account the framework imposed by the Council of Europe.

Online disinformation – a term preferred to fake news – is not *per se* illegal, although it may be harmful to society as it can be detrimental to the formation of informed and pluralistic opinions, which are essential for citizens to freely exercise their democratic choices. It can therefore be damaging to democratic elections, decreasing trust among citizens and creating tensions within society. The European Commission has defined online disinformation as "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm"[14]. Such an approach excludes unintentional journalistic errors. Moreover, the principle of the relationship to the truth ("verifiably false or misleading information") also excludes content that is part of the opinion's register. The European Commission also points out that it does not cover clearly identified partisan news and commentary.

2.2. EU regulatory framework on moderation of illegal content online

The EU regulatory framework on content moderation is complex as illustrated in Figure 1 below. It consists of:

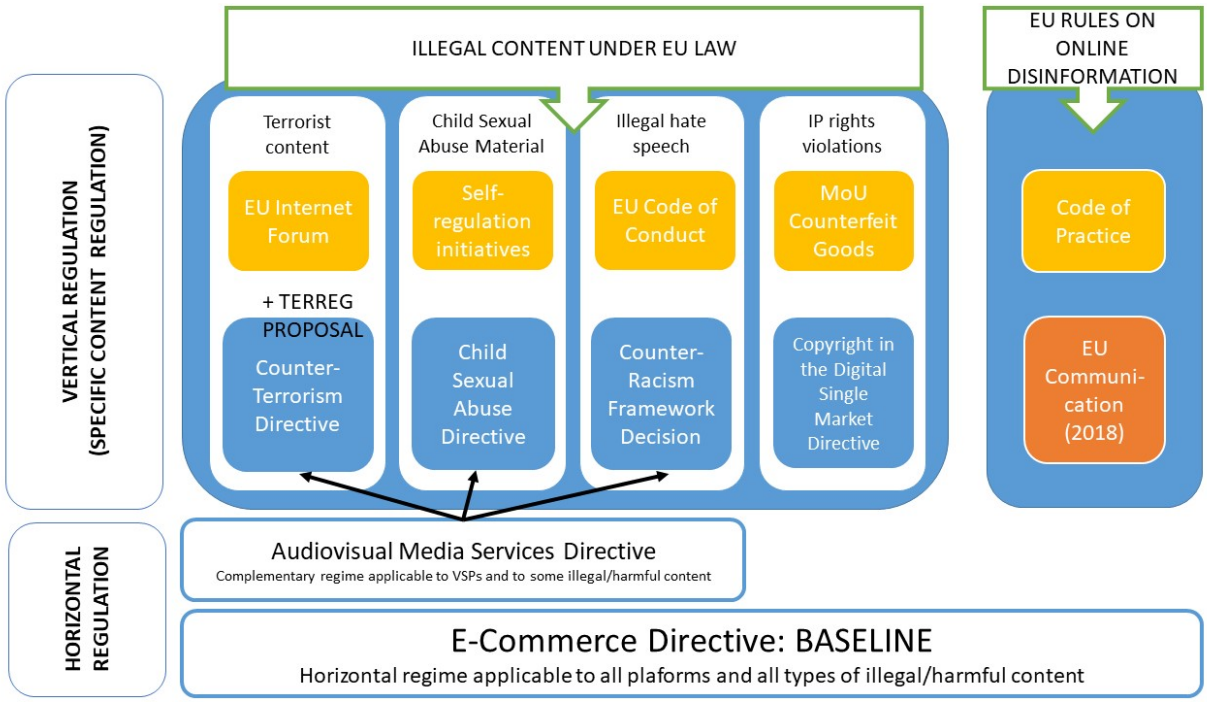
- some horizontal rules applicable to all categories of online platforms and to all types of content, i.e. the e-Commerce Directive (ECD)[15];
- some stricter rules applicable to Video-Sharing Platforms (VSPs) and to certain types of illegal content online, i.e. the revised Audio-visual Media Service Directive (AVMSD)[16]; and
- those rules are then complemented by vertical rules applicable to the four types of illegal content, which are illegal under EU law (i.e. terrorist content, child sexual abuse material, racist and xenophobic hate speech and violations of Intellectual Property).

[14] Communication from the European Commission of 26 April 2018, Tackling online disinformation: a European Approach, COM(2018) 236, pp. 3-4.

[15] Directive 2000/31 of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L 178/1.

[16] Directive 2010/13 of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audio-visual media services (Audio-Visual Media Services Directive), OJ [2010] L 95/1, as amended by Directive 2018/1808. The 2018 revision of the Directive which contains the new regime for VSPs should be transposed in the Member States by September 2020. On this Directive, see Valcke (2019).

Figure 1: EU regulatory framework for online content moderation



Source: Authors' own elaboration

2.2.1. EU rules applicable to all online platforms

a. Objective and scope

In 2000, when online platforms were in their infancy, the ECD established a special liability regime for online intermediary services. As explained by the European Commission[17], this legal regime pursued **four main objectives**:

- first, to share responsibility for a safe Internet between all the private actors involved and a good cooperation with public authorities, thus, injured parties should notify online platforms on any illegality they observe and online platforms should remove or block access to any illegal material of which they are aware;
- second, to encourage the development of e-Commerce in Europe by ensuring that the online platforms did not have an obligation to monitor the legality of all material they store;
- third, to strike a fair balance between different fundamental rights of the several stakeholders, in particular privacy and freedom of expression, freedom to conduct business (for platforms) and the right to property including Intellectual Property of injured parties[18]; and
- fourth, to strengthen the Digital Single Market by adopting a common EU standard on liability exemptions, especially at a time when national rules and case law were increasingly divergent.

[17] Commission, 'Explanatory Memorandum of the Commission proposal for a directive on certain legal aspects of electronic commerce in the internal market', COM(1998)586.

[18] As protected by the Charter of Fundamental Rights of the European Union, Article 7, 8, 11, 16 and 17.

While the ECD provides **liability exemptions for three categories of online platforms**, i.e. mere conduit, caching and hosting, this study focuses on the latter category which is defined as the storage of information provided by a recipient of the service[19].

Regarding the scope of the liability exemptions, reference should also be made to the criterion of **neutrality**. The Court of Justice decided that "in order to establish whether the liability of a referencing service provider may be limited under Article 14 of the ECD, it is necessary to examine whether the role played by that service provider is neutral, in the sense that its conduct is merely technical, automatic and passive, pointing to a lack of knowledge or control of the data which it stores"[20]. Following the case law of the Court of Justice, the exemption of liability cannot be excluded on the sole basis that the service is subject to payment, or that general information is provided by the platform to its clients. An assessment should be made by the national jurisdiction on a case-by-case basis and for that purpose, the special assistance provided by the online platform to the client could, for instance, be relevant[21].

A related issue is **whether the ECD dis-incentivises the online platforms to proactively monitor the legality of the material** they host because, if they would do so, they may lose the benefit of the liability exemption. This is sometimes referred to as the 'Good Samaritan' paradox[22]. For instance, a platform carrying out *ex ante* moderation practices could be considered as playing an active role and, therefore, be excluded from the liability exemption. During the public consultations that the European Commission did on the ECD, online platforms have mentioned this legal risk of voluntary introducing more proactive measures[23]. However, in its Communication of September 2017 on tackling illegal online content, the European Commission considers that voluntary proactive measures "do not in and of themselves lead to a loss of the liability exemption, in particular, the taking of such measures need not imply that the online platform concerned plays an active role which would no longer allow it to benefit from that exemption"[24]. Without any judgement of the Court of Justice or clarification of the legal framework, uncertainty remains.

---

19 ECD, Article 14. In 2000, the storage of information meant the provision of technical services consisting of storage space on a server (e.g. hosting an Internet website). With the so-called web 2.0 (social networks, electronic marketplaces, etc.), the question arose whether hosting also covered virtual storage of content provided by users. Virtual storage means that online platforms provide storage space for data provided by users (product sales announcements, messages or posts published on social networks, video content or photos exchanged on the platforms), so that such data are made available to other users through social networks or electronic marketplaces, etc. Although national judges sometimes decided that these kinds of services should not be considered as hosting services, the Court of Justice confirmed that they could fall within the scope of the liability exemption for hosting: Cases C-236/08 to C-238/08 *Google France v Louis Vuitton* EU:C:2010:159, para. 111; Case C-324/09 *L'Oreal et al. v. eBay* EU:C:2011:474, para. 110; Case C-360/10 *SABAM v. Netlog* EU:C:2012:85, para. 27.

20 Cases C-236/08 to C-238/08 *Google France v Louis Vuitton* EU:C:2010:159, para. 113 where the Court of Justice decided that: '*the exemptions from liability established in the directive cover only cases in which the activity of the information society service provider is 'of a mere technical, automatic and passive nature', which implies that that service provider 'has neither knowledge of nor control over the information which is transmitted or stored'*; Case C-324/09 *L'Oreal et al. v. eBay* EU:C:2011:474, para. 116 where the Court of Justice decided that: '*Where, the operator has provided assistance which entails, in particular, optimising the presentation of the offers for sale in question or promoting those offers, it must be considered not to have taken a neutral position between the customer-seller concerned and potential buyers but to have played an active role of such a kind as to give it knowledge of, or control over, the data relating to those offers for sale. It cannot then rely, in the case of those data, on the exemption from liability*'; Case C-484/14 *Mc Fadden* EU:C:2016:689, para. 62. Those cases are well explained in Van Eecke (2011), Husovec (2017), Nordemann (2018), Van Hoboken et al. (2018).

21 See he example of Google AdWords: should the provider play a role in the drafting of the commercial message which accompanies the advertising link: Cases C-236/08 to C-238/08, *Google France v Louis Vuitton,* EU:C:2010:159, para. 118.

22 Note that the US law provides explicitly for a 'Good Samaritan' clause, hence does not carry this dis-incentive against voluntary proactive measures: Section 230(c) of the US Communication Decency Act states that: '(…) *No provider or user of an interactive computer service shall be held liable on account of any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable* (…)'.

23 For the 2011 public consultation: Commission Staff Working Document of 11 January 2012, Online services, including e-Commerce, in the Single Market, SEC(2011) 1641, p.35. For the 2015-2016 consultation, Communication from the Commission of 25 May 2016, Online Platforms and the Digital Single Market Opportunities and Challenges for Europe, COM(2016) 288, p. 9 and Commission Staff Working Document of 10 May 2017 on the Mid-Term Review on the implementation of the Digital Single Market Strategy, SWD(2017) 155, p. 28.

24 Communication of the Commission of 28 September 2017, Tackling Illegal Content Online. Towards an enhanced responsibility for online platforms, COM (2017) 555, p.13.

b. Rights and duties of the hosting provider

Article 14 of the ECD creates an **exemption from the national liability regime** to which the hosting platform is subject and determine the requirements to be met by the providers to benefit from such exemption. Liability exemptions are horizontal: many types of illegal content or activities are covered (unfair market practices, violation of data protection rules, damage to honour and reputation, etc.), as well as various kinds of liabilities (criminal or civil). Note that, even when the platform cannot benefit from the liability exemption, it does not mean that it will necessarily be considered as liable under the applicable legal framework. In this case, the national jurisdiction should determine whether legal requirements applicable in the Member State are fulfilled (e.g. a negligence under Tort Law) and, if so, to decide that the online platform shall be held liable.

Following Article 14 of the ECD, a hosting platform can escape liability for illegal material uploaded by users when it "does not have actual knowledge of illegal activity or information and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or information is apparent". Should the platform have such knowledge or awareness, it can however benefit from the liability exemption if it "acts expeditiously to remove or to disable access to the information".

Since the adoption of the Directive in 2000, the interpretation of Article 14 gave rise to various discussions, in particular with regard to the concepts referred to in the provision ("*actual knowledge*", "*acting expeditiously*", etc.). In this context of legal uncertainty, the hosting provider has to manage conflicting claims between, on the one hand, the victim of the illegal content (who may engage the provider's liability if it is not removed or blocked) and, on the other hand, the originator of the content (whom the intermediary has contractually committed to host). Moreover, if for certain types of content the illicit character is obvious (child pornography, massive violation of copyright, etc.), for other types of content it is less certain (attack on honour or reputation, violation of image rights or privacy, etc.).

Another pillar of the ECD consists in the **prohibition for EU Member States to impose a general obligation on the hosting platforms to monitor** the material hosted [25]. The Court of Justice has drawn a blurred line between general monitoring measures and specific monitoring measures, in particular in case of suspected violation of Intellectual Property Rights. The first are prohibited [26]; the second are allowed when achieving a fair balance between the fundamental rights of the different stakeholders [27]. Although there is not any general obligation to monitor, the online platforms could decide, on a voluntary basis, to carry out spot checks on the online content. This is not prohibited, however, while doing so, the online platform could be considered as playing an active role (and therefore lose the benefit of the liability exemption).

---

[25] ECD, Article 15.

[26] Case C-360/10 *SABAM v. Netlog* EU:C:2012:85 where the Court of Justice decided that the e-Commerce Directive precludes: '*a national Court from issuing an injunction against a hosting service provider which requires it to install a system for filtering information which is stored on its servers by its service users; which applies indiscriminately to all of those users, as a preventative measure, exclusively at its expense, and for an unlimited period; which is capable of identifying electronic files containing musical, cinematographic or audiovisual work in respect of which the applicant for the injunction claims to hold intellectual property rights, with a view to preventing those works from being made available to the public in breach of copyright*'. Also Case C-70/10 *Scarlet Extended v. SABAM* EU:C:2011:771.

[27] Case C-314/12 *UPC Telekabel Wien v Constantin Film Verleih GmbH* EU:C:2014:192 where the Court of Justice decided that the injunction must: '*strike a balance, primarily, between (i) copyrights and related rights, which are intellectual property and are therefore protected under Article 17(2) of the Charter, (ii) the freedom to conduct a business, which economic agents such as internet service providers enjoy under Article 16 of the Charter, and (iii) the freedom of information of internet users, whose protection is ensured by Article 11 of the Charter*' (at para. 47 of the Case) and that such balance is found when the injunctions do not: '*unnecessarily deprive internet users of the possibility of lawfully accessing the information available and that they have the effect of preventing unauthorised access to protected subject-matter or, at least, of making it difficult to achieve and of seriously discouraging internet users who are using the services of the addressee of that injunction from accessing the subject-matter that has been made available to them in breach of the intellectual property right*' (at para. 63 of the case). Also more recently, Case C-484/14 *Mc Fadden*, para. 96.

In addition, Member States may impose on hosting providers the **duty to cooperate with the competent authorities**[28]. Two types of duties are possible: spontaneous communication to the authorities or communication at their request. Information related to the identification of the user who posted illegal content anonymously could be communicated to the victim of the illegal content (in order to bring a claim against the author) or only to the competent authorities.

The last pillar of the ECD is the **encouragement of co- and self-regulation** to implement the rules and principles of the Directive[29]. In particular, the ECD mentions the importance of involving consumers in drafting Codes of Conduct to ensure that the rules remain balanced. To ensure the effectiveness of those rules, monitoring the implementation of the codes is essential. This provision has led to an increasing reliance on co- and self-regulation to tackle certain types of illegal materials which have very negative impact on the society, such as child abuse content, terrorist content, hate speech or counterfeit goods.

c. Commission Recommendation of March 2018

In March 2018, the European Commission adopted a Recommendation setting principles for the providers of hosting services and Member States to take effective, appropriate and proportionate measures to tackle illegal content online[30]. It sets out the general principles for all types of illegal content online and recommended stricter moderation for terrorist content:

- regarding the **'notice-and-takedown'**, the Recommendation calls for procedures that (i) are effective, sufficiently precise and adequately substantiated, (ii) respect the rights of content providers with possibilities of 'counter-notices' and out-of-Court dispute settlements and (iii) are transparent[31];
- regarding **proactive measures**, the Recommendation encourages appropriate, proportionate and specific measures, which could involve the use of automated means, provided some safeguards be in place, in particular human oversight and verification[32]; and
- regarding **cooperation**, the Recommendation encourages close cooperation with national judicial and administrative authorities and trusted flaggers with the necessary expertise and determined on clear and objective basis; it also encourages cooperation among hosting services providers, in particular smaller ones which may have less capacity to tackle illegal content[33].

d. Evaluation of the rules

In 2016, the Commission did an evaluation of the ECD with a focus on the liability regime, its harmonisation across the EU and its effectiveness in tackling illegal content online, whose results should be taken into account in the forthcoming Digital Services Act. According to the 2016 Commission public consultation[34], "[a] **majority of the respondents stands behind intermediary liability principles** of the e-Commerce Directive, but also demands some clarifications or

[28] ECD, Article 15(2) requires to "promptly to inform the competent public authorities of alleged illegal activities undertaken or information provided by recipients of their service or obligations to communicate to the competent authorities, at their request, information enabling the identification of recipients of their service with whom they have storage agreements".
[29] ECD, Article 16.
[30] Recommendation 2018/334 of the European Commission of 1 March 2018 on measures to effectively tackle illegal content online, OJ [2018] L 63/50. This Recommendation follows the Communication of the European Commission of 28 September 2017, Tackling Illegal Content Online. Towards an enhanced responsibility for online platforms, COM (2017) 555.
[31] European Commission Recommendation 2018/334, Points 5-17.
[32] European Commission Recommendation 2018/334, Points 16-21.
[33] European Commission Recommendation 2018/334, Points 22-28.
[34] TILT (2016, p. 4).

improvements. A significant proportion of respondents who criticized the Directive complained about the national implementations rather than the EU law itself. The stakeholders broadly supported the horizontal nature of the Directive, but demanded a differentiated approach on 'Notice-and-takedown'; by adjusting or improving the practice of takedown for specific types of content, such as hate speech, terrorist content, child abuse material, copyright infringements, etc."

Regarding the functioning of the ECD rules, the most attention was paid to the hosting safe harbour, in particular its concept of **"passive hosting"**. **The concept was criticised for not being entirely clear, and for divergent national interpretations**. Regarding the missing components, an "[o]verwhelming majority of respondents supported the establishment of a 'counter-notice' mechanism (82.5%), i.e. possibility for content providers to give their views to the hosting service provider on the alleged illegality of their content"[35]. The consultation also recorded a significant support for more transparency on the intermediaries' content restriction policies[36]. On the side of proactive duties, a majority of intermediaries reported that they do put in place voluntary or proactive measures to remove certain categories of illegal content from their system beyond what was required by the legal framework. In the consultation, only 36.1% of respondents reported a need to impose specific duties of care for certain categories of content.

With regard to the liability of platforms, some of the empirical studies look at the question of removal of illegal content online. However, most of them are copyright-centred, and not necessarily focused on EU markets only. The other problem is that in the online environment, where many companies are operating globally with global version of products, the broadly drafted EU rules sometimes could give a way to more prescriptive rules from the US (especially in copyright law).

As noted by de Streel and Husovec (2020), among the academic studies, the studies of the ecosystem fit into several categories: (i) interviewing notifiers, providers and users[37]; (ii) experimental upload of content[38]; (iii) analysis of transparency reports or data sets shared publicly by providers, such as *Lumen* data[39]; (iv) tracking of the public availability of the content over a pre-set period[40]; and (v) experimental testing of redesign of the ECD[41]. The studies so far show a number of global trends, which are not always restricted to the European setting, namely:

- the **quality of notifications** sent to the providers is often low (at least in some areas) and there is a diverging quality of such notifications among different actors;
- the notifications are **increasingly out-sourced** to professional companies and also sent by **algorithms**, and not humans; and
- providers tend to **over-remove content** to avoid liability and save resources, they equally employ technology to evaluate the notifications; and the affected users who posted content often do not take action.

35 *Ibidem*, p. 5.
36 *Ibidem*, p. 6.
37 Urban et al. (2017a).
38 Perel and Elkin-Koren (2017); Sjoera (2004).
39 See www.lumendatabase.org; Urban and Quilter (2006); Urban et al. (2017a) and (2017b); Seng (2014) and (2015).
40 Erickson and Kretschmer (2018).
41 Fiala and Husovec (2018).

e. EU data protection rules to protect privacy

Two key EU legislations, which **regulate the collection and the processing of personal data** in order to protect the privacy of EU citizens, have important **impacts on the online content moderation practices**:

- first, the General Data Protection Regulation (GDPR) imposes strict rules for the collection and the processing of personal data, in particular the principles of lawfulness/fairness/transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity/confidentiality and accountability[42]. Thus any content moderation practices involving personal data, which is often the case given the broad definition of personal data, should comply with those principles; and
- second, the e-Privacy Directive complements the GDPR and imposes even stricter rules for telecommunications operators, in particular to protect the confidentiality of communications[43]. In 2017, the European Commission proposed a new e-Privacy Regulation in order to update the current legislation and to better align it with the new rules of the GDPR. However, the proposal is still being discussed in the Council, and negotiations between the two EU co-legislators have not yet started.

2.2.2. Additional rules applicable to Video-Sharing Platforms

The 2018 revision of the AVMSD imposes on the VSPs[44] to take appropriate measures to[45]:

- protect **the general public** from: (i) the **three types of online content which are illegal under EU law** (terrorist content, child sexual abuse material and racism and xenophobia) and (ii) **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)[46]; and
- protect **minors** from **content which may impair their physical, mental or moral development**[47].

The AVMSD lists the possible measures to be taken such as transparent and user-friendly mechanisms to report and flag the content; systems through which VSPs explain to users what effect has been given

---

42 Regulation 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 (General Data Protection Regulation), OJ [2016] L 199/1, Article 5.

43 Directive 2002/58 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), OJ [2002] L 201/37 as amended by Directive 2009/136. The European Commission has proposed a new e-Privacy Regulation in order to update the legislation and better align it with the new rules of the GDPR. However, the proposal is still being discussed in the Council, and negotiations between the EU co-legislators have not yet started: Proposal of the Commission of 10 January 2017 for Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58 (Regulation on Privacy and Electronic Communications), COM(2017) 10.

44 AVMSD, Article 1(1aa) defines the Video-Sharing Platform service as "a service as defined by Articles 56 and 57 TFEU, where the principal purpose of the service or of a dissociable section thereof or an essential functionality of the service is devoted to providing programmes, user-generated videos, or both, to the general public, for which the Video-Sharing Platform provider does not have editorial responsibility, in order to inform, entertain or educate, by means of electronic communications networks (…) and the organisation of which is determined by the Video-Sharing Platform provider, including by automatic means or algorithms in particular by displaying, tagging and sequencing".

45 AVMSD, Article 28b. As AVMSD is based on the minimum harmonisation, Member States can adopt more detailed or stricter rules provided they respect EU law, in particular the ECD and the CSAED.

46 AVMSD, Article 28b (1b) and (1c).

47 AVMSD, Article 28b (1a); content in the meaning of the AVMSD, i.e. programmes, user-generated videos and audio-visual commercial communications.

to the reporting and flagging; easy-to-use systems allowing users to rate the content; transparent, easy-to-use and effective procedures for the handling and resolution of users' complaints[48]. However, as explained by Kukliš (2020), the AVMSD "does not create a duty of care or any other general responsibility of VSPs vis-à-vis their users, nor does it create any actual substantive rights worth that name."

The Directive specifies that the measures must be **appropriate** in the light of the nature of the content, the potential harm, the characteristics of the category of persons to be protected, the rights and legitimate interests at stake (in particular those of the VSPs and the users having created and/or uploaded the content, as well as the public interest). The measures should also be **proportionate** taking into account the size of the VSP and the nature of the provided service[49]. A National Regulatory Authority (often the media regulator) should assess the appropriateness of the measures[50]. According to the European Commission, the requirements of the AVMSD are compatible with the liability exemption for hosting service providers of the ECD, as the measures imposed on VSPs relate to the responsibilities of the provider in the organisational sphere and do not entail liability for any illegal information stored on the online platforms as such[51]. Moreover, the measures imposed on VSPs cannot lead to any *ex-ante* control measures or upload-filtering of content[52].

In its Impact Assessment leading to the proposal to amend the AVMSD[53], the Commission considered that minors and consumers were not sufficiently protected when viewing videos on VSPs. By encouraging co-regulatory measures, the additional cost would be limited because most online platforms have already similar mechanisms in place and the costs are shared between the industry and regulators. Moreover, the proposed measures strike an adequate balance between, on the one hand, the need to enhance the protection viewers and, on the other hand, the need to protect the fundamental rights (in particular, the freedom of speech and the freedom to conduct a business).

2.2.3. Stricter rules applicable for terrorist content

To better fight terrorist content online, the baseline regulatory regime is complemented by the following legislative and non-legislative elements:

- the CTD, which may be complemented by TERREG if adopted by the co-legislators;
- the Guidelines of the European Commission on effectively tackling illegal content online[54]; and
- a Forum with the main stakeholders involved in the fight against terrorist content online, which has been established in 2015.

a. Counter-Terrorism Directive

The CTD requires **Member States to take content removal and blocking measures against websites containing or disseminating terrorist content**[55]. Those measures must be set following transparent procedures and provide adequate safeguards, in particular to ensure that they are limited to what is necessary and proportionate and that users are informed of the reason for those measures.

---

48 AVMSD, Article 28b (3).
49 AVMSD, Article 28b (3).
50 AVMSD, Article 28b (5).
51 Proposal for a Directive of the European Parliament and of the Council 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 audio-visual media services in view of changing market realities, COM(2016) 287.
52 AVMSD, Article 28b(3).
53 Commission Staff Working of 25 May 2016, Impact Assessment of AVMSD Proposal, SWD(2016) 168.
54 European Commission Recommendation 2018/334, Chapter III.
55 CTD, Article 21.

As of September 2018, 15 Member States had already transposed the Directive by adopting two main types of measures[56]:

- The **'notice-and-takedown' measures** under the ECD, which differ on several points among the Member States: offences covered, time limits for removal and consequences of non-compliance; and
- The **criminal law measures** allowing a prosecutor or a Court to order companies to remove content or block content or a website, within a period of 24 or 48 hours in some circumstances.

However, those measures may not lead to a sufficient reduction of terrorist content online because it does not target directly the online platforms which are in the best position to know what technology can be applied and how to ensure a safe online environment for their users. Furthermore, those measures do lead to an effective EU-wide approach to ensure the removal of terrorist content with proactive efforts[57].

b. Proposal for a Regulation on preventing the dissemination of terrorist content online

The TERREG Proposal goes further than the CTD as it **imposes duties of care and proactive measures on Hosting Services Providers (HSPs) to remove terrorist content**[58]. The main elements of the proposal are:

- removal orders within one hour issued by a national competent authority, not necessarily a judicial body[59];
- content referrals sent from either a national competent authority or an EU body such as Europol that the HSPs must expeditiously assess[60]; and
- proactive measures taken by HSPs – when appropriate – to remove terrorist material from their services, including by deploying automated detection tools[61].

This system (and in particular the proactive measures) has been criticised by Van Hoboken (2019) as **undermining the ECD safe harbour**, by "creating a proactive duty of care for hosting service providers and moving beyond the reactive notice-and-takedown obligations that follow from the ECD framework"[62]. Moreover, such a system carries the risk of **negatively affecting fundamental rights**, in particular the right to freedom of expression:

- the one-hour response deadline is very difficult to meet in practice. Moreover, small online platforms are unlikely to have the resources to comply with this obligation. 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"[63]. According to Van Hoboken

---

56 Commission Staff Working Document of 12 September 2018, Impact Assessment TERREG Proposal, SWD(2018) 408, p. 22.
57 *Ibidem*, pp. 18-22.
58 TERREG Proposal, Article 2(1) defines Hosting Service Provider as "a provider of information society services consisting in the storage of information provided by and at the request of the content provider and in making the information stored available to third parties." HSPs are e.g. social media, Video-Sharing Platforms (VSP), cloud services and websites where users can post comments.
59 TERREG Proposal, Article 4.
60 TERREG Proposal, Article 5. Article 2(8) defines "referral" as "a notice by a competent authority or, where applicable, a relevant Union body to a hosting service provider about information that may be considered terrorist content, for the provider's voluntary consideration of the compatibility with its own terms and conditions aimed to prevent dissemination of terrorism content".
61 TERREG Proposal, Article 6.
62 Also in this sense, Kuczerawy (2018).
63 *Article 19* and Others (2018), *Joint letter on European Commission regulation on online terrorist content*, available at: https://www.article19.org/resources/joint-letter-on-european-commission-regulation-on-online-terrorist-content/.

(2019), a more flexible obligation to act quickly without undue delay would better support the necessary and proportionate requirement of interference with speech online;

- the proposal does not impose an independent judicial review for takedown orders. Moreover, the proposal does not provide the content provider or the HSP with a mechanism to effectively challenge the order before the removal is executed. The European Union Agency for Fundamental Rights (FRA, 2019) recommends that the competent authority issuing removal orders should be an independent judicial authority or to guarantee minimum standards concerning the reviewing of the removal orders;
- the proposal does not include sufficient safeguards for affected speakers/audiences. Not only service providers, but also other users can rely on the freedom of expression when confronted with takedown orders and referrals (e.g. Internet users who would have wanted to access the deleted content). Van Hoboken (2019) suggests "broadening standing in relevant appeal procedures beyond the user (re-)posting particular content to others unduly impacted in their freedom of expression";
- proactive measures require automated means, which may threaten the freedom of expression. Indeed, these means do not include safeguards to prevent abuse or provide redress when content is mistakenly removed. Furthermore, the proposal does not provide for appropriate transparency, accountability and redress mechanisms to mitigate this threat; and
- finally, this obligation applies to all hosting service providers, irrespective of their size, scope, purpose or revenue models, and does not allow flexibility for collaborative platforms[64].

In April 2019, the **European Parliament made several amendments to the Commission proposal**[65]. To ensure that the competent authority should be a judicial or a functionally independent administrative authority[66], to remove the provision concerning the referral[67], and to increase the safeguards to fundamental rights (including a ban on general monitoring, remedies and complaints mechanisms, transparency obligations on HSPs)[68].

c. EU Internet Forum

In December 2015, **the EU Internet Forum to counter terrorist content online was established among EU Interior Ministers, high-level representatives of major online platforms** (such as Facebook, Google, Microsoft and Twitter), Europol, the EU Counter-Terrorism Coordinator and the European Parliament[69]. One of its goals is to address the misuse of the Internet by terrorist groups and to reduce accessibility to terrorist content online.

The Forum led to an efficient referral mechanism in particular with the EU Internet Referral Unit of Europol, a shared database with more than 200,000 hashes, which are unique digital fingerprints of terrorist videos and images removed from online platforms. At its third meeting in December 2017, online platforms noted the increasing use and accuracy of Artificial Intelligence (AI), such as photo and video matching and text-based machine learning to identify terrorist content[70]. At its fourth meeting

---

64 *Ibidem*.
65 LIBE Committee Report of April 2019 on the proposal for a regulation on preventing the dissemination of terrorist content online (C8-0405/2018).
66 *Ibidem*, AM n° 126.
67 *Ibidem*, AM n° 83.
68 For example, AM n° 100, 103, 104, 106, 129.
69 European Commission Press release of 3 December 2015, IP/15/6243.
70 European Commission Press release of 6 December 2017, IP/17/5105. For instance, the Google *representative notes that* "98 percent of the videos we remove for violent extremism on YouTube are flagged to us by machine-learning algorithms, up from 75 percent just a few months ago".

in December 2018, participants stressed the importance of cooperation between public and private sectors and noted that out of more than 77,000 reported contents, 84% have been removed from online platforms[71]. During its fifth meeting in October 2019, participants committed to setting up an EU crisis protocol between the European Commission and Europol to facilitate international cooperation in the event of extraordinary situations for which national legal frameworks and crisis management mechanism are insufficient[72].

2.2.4. Stricter rules applicable for child sexual abuse material

To better fight child sexual abuse online material, a Directive against child sexual abuse was adopted in 2011 complemented by several self-regulatory initiatives.

a. Child Sexual Exploitation Directive

Similarly to the CTD, the CSAED requires **Member States to take content removal and blocking measures against websites containing or disseminating child sexual abuse material**. Such measures must be based on transparent procedures and provide adequate safeguards, in particular be necessary and proportionate, inform the users on the reasons for restriction and ensure the possibility of judicial redress[73].

To ensure the prompt removal of web pages containing or disseminating child pornography, Member States have adopted two categories of measures[74]:

- first, **'notice-and-takedown' measures** based on the ECD with national hotlines to which Internet users can report child sexual abuse material that they find online[75]; Moreover, INHOPE, a global umbrella organisation for the hotlines, encourages exchange of expertise[76]; and
- second, **measures based on national criminal law** such as general provisions that allow the seizure of material relevant to criminal proceedings (e.g. material used in the commission of an offence) or more specific provisions on the removal of child sexual abuse material.

With regard to the optional blocking measures, about half of the Member States have chosen to apply such measures by using various means (legislative, non-legislative, judicial or other, including voluntary action by the Internet industry). Blacklists of websites containing or disseminating child sexual abuse material are commonly used in the implementation of blocking measures.

Noting that "in the fight against the dissemination of child sexual abuse material, removal measures are more effective than blocking, since the latter does not delete the content", the **European Parliament recommends further measures**, such as the speeding up of 'notice-and-takedown' procedures in cooperation with the Internet industry, the removal of child sexual abuse material at source with efficient judicial and law enforcement actions, the establishment of partnerships (with online platforms, Europol and Eurojust) to prevent networks and systems from being hacked and misused to distribute child sexual abuse material, the legal obligation for Internet Service Providers

---

71 European Commission Statement of 5 December 2018, Statement/18/6681.

72 European Commission Press release of 7 October 2019, IP/19/6009.

73 CSAED, Article 25. Measures may consist in various types of public action, such as legislative, non-legislative, judicial or others.

74 Report from the Commission of 16 December 2016 assessing the implementation of the measures referred to in Article 25 of Directive 2011/93 on combating the sexual abuse and sexual exploitation of children and child pornography, COM(2016) 872.

75 In practice, a person can anonymously report illegal content online to a hotline. Content analysts review the reported content, classify the illegality of the material and warn the local law enforcement agency. In many cases, the relevant Internet Service Provider will receive a 'notice-and-takedown' order.

76 This organisation is supported by the European Commission's Safer Internet Programme and since 2014, by the Connecting Europe Facility framework. The Network consists of 47 hotlines in 43 countries (as of June 2019). Hotlines have memoranda of understanding with the corresponding national Law Enforcement Agencies. See INHOPE Annual Report 2018.

(ISPs) to report child sexual abuse material detected in their infrastructure proactively to law enforcement authorities and national hotlines[77].

b. Alliance to Better Protect Minors Online

In 2017, the **Alliance to Better Protect Minors Online, a multi-stakeholder forum facilitated by the European Commission, was set up in order to address emerging risks that minors face online, such as illegal and harmful content** (e.g. violent or sexually exploitative content), conduct (e.g. cyberbullying) and contact (e.g. sexual extortion)[78]. It is composed of actors from the entire value chain (device manufacturers, telecoms, media and online platforms used by children)[79]. Its action plan includes the provision of accessible and robust tools that are easy-to-use, the provision of feedback and notification as appropriate, the promotion of content classification when and where appropriate, and the strengthening of the cooperation between the members of the Alliance and other parties (such as Child Safety Organisations, Governments, education services and law enforcement) to enhance best practice-sharing[80].

In its evaluation, Ramboll (2018) indicates that **many commitments are difficult to measure**, hence their effectiveness is difficult to assess. It also notes that the effectiveness of the Alliance is limited by **low public awareness and limited internal knowledge-sharing**. It therefore recommends to increase public awareness in order to strengthen the external monitoring of the commitments and to incentivise the Alliance participants to meet them and to reinforce sharing of good practices between members. It also recommends to intensify discussions on new technologies and that they become more central to the commitments.

2.2.5. Stricter rules applicable for racist and xenophobic hate speech

To better fight racist and xenophobic hate speech, the Counter-Racism Framework Decision was adopted in 2008 but does not deal specifically with online content and a Code of Conduct has been agreed in 2016.

a. Counter-Racism Framework Decision

The **CRFD provides that Member States must ensure that hate speech is punishable but does not provide for detailed obligations related to online content moderation practices**, contrary to the CTD or the CSAED. In its Implementation Report, the European Commission indicates that the fragmentation of criminal procedural rules across Member States make it difficult to enforce the

---

77 European Parliament Resolution of 14 December 2017 on the implementation of Directive 2011/93 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)), in particular, para 40-53.

78 European Commission, "Alliance to better protect minors online", available at: https://ec.europa.eu/digital-single-market/en/alliance-better-protect-minors-online. Previous self-regulatory initiatives were: The a CEO Coalition to Make the Internet a Better place for Kids set up in 2011 (https://ec.europa.eu/digital-single-market/en/self-regulation-and-stakeholders-better-internet-kids) and the ICT Coalition for Children Online set up in 2012 (http://www.ictcoalition.eu).

79 The signatory companies are: ASKfm, BT Group, Deutsche Telekom, Disney, Facebook, Google, KPN, The LEGO Group, Liberty Global, Microsoft, Orange, Rovio, Samsung Electronics, Sky, Snap, Spotify, Sulake, Super RTL/Mediengruppe RTL Deutschland, TIM (Telecom Italia), Telefónica, Telenor, Telia Company, Twitter, Vivendi, Vodafone. The associated signatories are: BBFC, Child Helpline International, COFACE, eNACSO, EUN Partnership, FFTelecoms, FOSI, Foundation T.I.M., FSM, GSMA, ICT Coalition, NICAM, Toy Industries of Europe, UNICEF. However, the independent evaluation of the implementation of the Alliance shows that the majority of members are for now large companies and the telecommunication sector. The report therefore calls for more representation of small businesses and for a better diversity in the range of stakeholders to increase mutual learning opportunities: Ramboll, 2018.

80 The common action is complemented by individual company commitments with specific timeline to better protect minors online, see: https://ec.europa.eu/digital-single-market/en/news/individual-company-statements-alliance-better-protect-minors-online.

Framework Decision and many Member States could not provide detailed information on how their general jurisdictional rules cover online hate speech situations[81].

b. EU Code of Conduct on countering illegal hate speech online

In 2016, the **main online platforms agreed, at the initiative of the European Commission, an EU Code of Conduct on countering illegal hate speech online[82] and commit to fight the dissemination of illegal hate speech as defined according in the CRT**. The Code considers that online platforms have a role key role to play in ensuring compliance with CRT and the platforms have made a series of commitments:

- drawing users' attention to the types of content not allowed by their Community Standards/Guidelines and specifying that they prohibit the promotion of incitement to violence and hateful behaviour;
- putting in place a clear and effective process to review reports/notifications of illegal hate speech to remove them or make them inaccessible; reviewing notifications on the basis of the Community Standards/Guidelines and the national transposition laws, and reviewing the majority of valid reports within 24 hours;
- regularly training online platform staff, particularly in relation to societal developments;
- encouraging the reporting of illegal hate speech by experts, including through partnerships with Civil Society Organisations (CSOs) - so that they can potentially act as trusted reporters - and strengthening partnerships and collaboration with CSOs to support them; and
- strengthening communication and cooperation between the online platforms and the national authorities, in particular with regard to procedures for submitting notifications; collaborating with other online platforms to improve and ensure the exchange of best practices between them.

The implementation and the impact of the Code is regularly assessed by the Commission on the basis of information given by the platforms. The fourth evaluation of January 2019 indicates that[83]: **88.9% of notifications are reviewed within 24 hours** (up to 40% in 2016), **the speed of the review of notifications improves and an average of 71.7% of reported illegal hate speech is removed**. However, there is little information on how the statistics are calculated. With regard to transparency towards users, only 65.4% of notifications receive feedback and only Facebook systematically informs users by providing feedback[84]. The percentage of feedback is higher when the notification came from a trusted flagger[85]. Quintel and Ullrich (2019) note that the evaluation of effectiveness focuses on the number and speed of removals but not on the actual illegality of the removed content.

---

81 Report of the European Commission of 27 January 2014 on the implementation of Council Framework Decision 2008/913 on combating certain forms and expressions of racism and xenophobia by means of criminal law, COM(2014)27.

82 The Code of Conduct was signed in 2016 by Facebook, Microsoft, Twitter and YouTube. Since then, Google+, Instagram, Dailymotion and Snapchat and Jeuxvideo.com have joined. The Code is available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

83 European Commission, Code of Conduct on countering illegal hate speech online: fourth evaluation confirms self-regulation works, Factsheet, February 2019, available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#theeucodeofconduct.

84 Facebook 92.6%, Twitter 60.4% and YouTube only 24.6%.

85 Facebook 96.8%, Twitter 88.2%, Youtube 40.5% and Instagram 95.5%.

Commentators have pointed towards the following **weaknesses**[86]:

- risks of private censorship practices through the priority application of Community Standards/Guidelines;
- lack of precision in determining the validity of a notification;
- absence of appeal mechanisms for users whose content has been withdrawn;
- illegal content does not have to be reported to the competent national authorities when removed on the basis of the Community Standards/Guidelines; and
- the 24-hour deadline could either make it impossible for online platforms to meet their commitments or lead them to over-blocking practices.

2.2.6. Stricter rules applicable for violation of Intellectual Property

a. Copyright in the Digital Single Market Directive

The CDSMD states that Online Content-Sharing Service Providers perform an act of communication to the public or an act of making available to the public when they give the public access to protected works or other protected subject matter which are uploaded by their users[87]. Therefore, the Directive provides the platforms with this option to avoid direct liability for their users' uploads[88]:

- conclude an agreement with the right holder for the exploitation of the works; or
- (i) make their best efforts to obtain an authorisation from copyright holders, (ii) make their **best efforts to ensure the unavailability of specific works violating copyright**, which the rights holders have provided them with the relevant and necessary information and (iii) act expeditiously to disable access to, or to remove from their websites, the notified works, and make their best efforts to prevent their future uploads.

The CDSMD also provides for **user safeguards to minimise the risks of broad filtering and over-blocking**[89]. Internet service providers have to put in place rapid and effective measures to enable users of their services to lodge a complaint against the blocking or removal of content. Complaints shall be processed without undue delay, and decisions to disable access to or remove uploaded content shall be subject to human review (not an automated device).

b. Memorandum of Understanding on the sale of counterfeit goods via the Internet

A **Memorandum of Understanding on the sale of counterfeit goods via the Internet was signed in 2011 between rights owners, online platforms and associations**[90]. It aims to **improve 'notice-and-takedown' measures and enhance proactive measures** taken by rights owners and online intermediaries, to increase cooperation and to better fight against repeated infringements. In its first evaluation report, the European Commission noted that voluntary cooperation has been a useful tool to reduce online counterfeiting and piracy when used alongside legislation and offered the flexibility

---

86 Coche (2018); Quintel and Ullrich (2019).
87 Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9 and 2001/29, OJ [2019] L 130/92, Article 17(1).
88 CDSMD, Article 17(4).
89 CDSMD, Article 17(9).
90 Memorandum of Understanding of 21 June 2016 on the sale of counterfeit goods via the Internet, available at: https://ec.europa.eu/growth/industry/policy/intellectual-property/enforcement/memorandum-understanding-sale-counterfeit-goods-internet_en.

to quickly adapt to technological developments and deliver efficient solutions[91]. However, it also noted the need to measure more precisely the effects of the Memorandum of Understanding (MoU).

Accordingly, **a set of Key Performance Indicators (KPIs) has been added in a revised version of the MoU in 2016**[92] and data based on KPIs are collected every six months. In its overview of the revised MoU, the Commission indicates that the number of lists deleted as a result of the commitments taken by online platforms increases almost tenfold between December 2016 and June 2017 (from 2.65 to 13.7%)[93]. Moreover, the feedback received shows that 'notice-and-takedown' measures are useful and have been improved by the MoU. However, as they are only *ex-post*, they need to be complemented by proactive measures. Those measures require a close cooperation between online platforms and right holders.

2.2.7. Summary of the EU regulatory framework

Table 1 outlines the EU rules against illegal content online according to the nature of the legal instrument (hard-law, soft-law, or self-regulation). The baseline regime is contained in the e-Commerce Directive, which applies to all categories of hosting platforms and all types of illegal content online. The Audio-Visual Media Services Directive provides for additional rules applicable to Video-Sharing Platforms and to certain type of illegal content online. In addition, a number of vertical measures have been adopted, applicable to specific type of content (terrorist content, child sexual abuse material, racist and xenophobic hate speech, intellectual property violations).

[91] Report of the European Commission of 18 April 2013 on the functioning of the Memorandum of Understanding on the Sale of Counterfeit Goods via the Internet, COM(2013) 209.

[92] Memorandum of Understanding of 21 June 2016 on the sale of counterfeit goods via the Internet.

[93] European Commission Staff Working Document of 29 November 2017, Overview of the functioning of the Memorandum of Understanding on the sale of counterfeit goods via the internet, SWD(2017) 430, p. 4.

Table 1: Main EU rules against illegal content online

Type of illegal content	Hard-law	Soft-law	Self-regulation
BASELINE ***All types of hosting platforms and all types of illegal content online***	- Directive 2000/31 on e-Commerce.	- Communication (2017) on Tackling *Illegal Content Online.* - Commission *Recommendation 2018/334* on measures to effectively tackle illegal content online.	
Additional rules for Video-Sharing Platforms	- Directive 2010/13 Audio-Visual Media Services as amended by Directive 2018/1808.		
Terrorist content	- Directive 2017/541 on combating Terrorism. - Proposal Regulation on preventing the dissemination of preventing the dissemination of terrorist content online.	- Commission *Recommendation 2018/334* on measures to effectively tackle illegal content online.	- EU Internet Forum (2015).
Child sexual abuse material	- Directive 2011/93 on combating the sexual abuse and sexual exploitation of children and child pornography.		- Alliance to Better Protect Minors Online (2017).
Illegal hate speech	- Council Framework Decision 2008/913 on combating certain forms and expressions of racism and xenophobia.		- Code of Conduct on illegal hate speech online (2016).
Intellectual property violation	- Directive 2019/790 on Copyright in the Digital Single Market. - Directive 2004/48 on enforcement of Intellectual Property Rights.		- Memorandum of Understanding on counterfeit goods online (2011, rev. 2016).

Source: Authors' own elaboration

Table 2 compares the main **elements of the EU legislations on online content moderation** on the basis of the following criteria: the type of online content, the category of online platforms, and the obligations imposed.

Table 2: Comparing EU legislations on online content moderation

	E-Commerce Directive	Audio-Visual Media Services Directive	Child sexual Exploitation and Abuse Directive	Counter-Terrorism Directive	Directive on Copyright in the Digital Single Market
Type of illegal content	All.	- Public provocation to commit a terrorist offence. - Child sexual abuse material. - Racist and xenophobic hate speech.	Child sexual abuse material.	Online content constituting a public provocation to commit a terrorist offence.	Content in breach of copyright and related rights.
Target	Hosting platforms.	Video-Sharing Platform Services.	Member States.	Member States.	Online Content-Sharing Service Providers.
Obligations	(Liability exemption).	Procedural accountability.	Blocking and removal measures.	Blocking and removal measures.	Liability exemption if best efforts.

Source: Authors' own elaboration

2.3. EU rules regarding the moderation of online disinformation

a. Communication from the European Commission

On the basis of a Report of the High-Level Expert Group on Fake News and Online Disinformation[94], the results of the public consultation and a Eurobarometer survey[95], the European Commission adopted in April 2018 a Communication on tackling online disinformation which is based on four principles[96]:

- **credibility of information** with trustworthiness indicators, trusted flaggers and information traceability measures; to guarantee their credibility, fact-checkers must maintain their independence and comply with strict rules of ethics and transparency, such as the International Fact-Checking Network's Code of Principles;

[94] A High-Level Expert Group (HLEG) on Fake News and Online Disinformation was set up by the European Commission. In its report, the HLEG considers that most effective solutions to combat online disinformation, while respecting freedom of expression, must involve the collaboration of all stakeholders and should be based on five pillars: transparency, media literacy, development of technical tools, preservation of media diversity and sustainability and continuous scientific research and evaluation of measures: High-Level Expert Group on Fake News and Online Disinformation (2018).

[95] Synopsis Report of the European Commission of 26 April 2018 of the public consultation on fake news and online disinformation, available at: https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation. Report of 27 April 2018 on Fake news and disinformation online Flash Eurobarometer 464, available at: https://publications.europa.eu/en/publication-detail/-/publication/2d79b85a-4cea-11e8-be1d-01aa75ed71a1/language-en.

[96] Communication from the European Commission of 26 April 2018, Tackling online disinformation: a European approach, COM(2018) 236.

- **transparency** of the origin, production, dissemination, targeting and sponsorship of information (especially for political advertisings and sponsored contents) to enable Internet users to directly evaluate online information and to detect any disinformation;
- **diversified offer of information** to citizens to encourage informed decisions based on critical thinking; this requires support for media literacy and quality journalism that play an important role in uncovering, counter-balancing, and diluting disinformation; and
- **inclusive solutions**, including awareness-raising campaigns for Internet users, media literacy and the mobilisation and the cooperation of all stakeholders (public authorities, online platforms, advertisers, fact and source checkers, journalists and media, etc.).

b. Code of Practice on Disinformation

In September 2018, some of the biggest online platforms (Facebook, Google, Twitter, Mozilla and Microsoft) and advertisers, as well as the advertising industry agreed a **Code of Practice on Disinformation with several commitments**[97]. **Some of the commitments directly relate 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; or facilitating the finding of alternative content on issues of general interest.

Other commitments can support better content moderation practices, such as: improving transparency of political and issue-based advertising, in particular by clearly distinguishing advertising content from editorial content and by disclosing the identity of the sponsor and the amounts spent; refusing remuneration and placements from accounts or websites that consistently disseminate disinformation; empowering the research community, fact-checkers and other relevant stakeholders, for instance with better access to data; or ensuring partnerships with other stakeholders to improve critical thinking and media literacy.

The European Commission monitored the actions undertaken by the stakeholders on a monthly basis before the European elections (between January and May 2019). In October 2019, the first annual Code of Practice self-assessment reports by the stakeholders were adopted. On that basis, the Commission notes **an improved transparency and closer dialogue on policies against online disinformation**. **However, the progress achieved differs greatly between online platforms, and the self-assessment reports do not provide sufficient information** on the impact of the measures undertaken and on mechanisms for independent scrutiny[98].

In its assessment Report, the European Regulators Group for Audio-Visual Media Services (ERGA, 2020) notes (i) a need for greater transparency on the implementation of the Code with a mechanism to ensure independent verification of information provided; (ii) the overly general nature of the commitments (both in terms of content and structure); and (iii) the need to increase the number of signatories, in particular to include all the big platforms. ERGA believes that improving the effectiveness of the Code requires that all online platforms must uniformly comply with the same obligations and that more precise definitions, procedures and commitments need be adopted. ERGA calls for a shift

[97] Code of Practice on Disinformation", 26 September 2018, last updated 17 June 2019, available at: https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

[98] European Commission Statement of 29 October 2019, Statement/19/6166.

from self-regulation to co-regulation to enhance the effectiveness of the fight against online disinformation.

In an Evaluation Study done for the European Commission, VVA (2020) analyses the Terms of Service/Use and Community Standards/Guidelines that online platforms have implemented to comply with the Code of Practice. The study makes three main criticisms: (i) given its self-regulatory nature, it is not possible to force signatories to comply with their commitments and they do not cover all stakeholders; (ii) fragmented implementation of the commitments across the different online platforms, pillars and Member States; and (iii) a lack of clarity around the scope and the key concepts of the Code of Practice. In this respect, VVA suggests, on the one hand, the adoption of a common terminology among signatories and, on the other hand, that the actions undertaken should be as concrete as possible. This would make it easier to implement and monitor the commitments and to define expected results and key performance indicators. VVA also makes recommendation to strengthen the effectiveness of the Code of Practice such more debates on the strengths and weaknesses of the Code, establishing mechanism for sanctions and redress in case of non-compliance with the commitments in the Code while considering proposals for co-regulation.

2.4. Summary of some national laws and initiatives on online content moderation

Next to the EU regulatory framework, national laws impose, for some categories of online platforms, additional obligations to moderate some types of illegal content online. This study focuses on Germany, France and the UK as their respective laws or proposals have been heavily debated in the policy circles and the academic literature. In Germany, the Network Enforcement Act (NetzDG) was adopted in June 2017 to improve the enforcement of existing criminal provisions on the Internet and, more specifically, on social networks[99]. In France, two related laws on information manipulation were adopted in December 2018 and a law on online hate speech, the so-called Avia law, was adopted in May 2020[100]. The legal compatibility of those national initiatives with the EU legal framework is not always clear. Moreover, the multiplication of national laws seriously risks undermining the Digital Single Market. In the UK, the Online Harms White Paper with proposals to combat online harms was adopted in April 2019[101].

Table 3 below summarises the main elements of the national laws and initiatives, which are analysed in more details in Annex I of this study. The main points emerging from the comparison are the following:

- regarding the **category of online platform**, the online platform must have an activity that exceeds a certain threshold. The German NetzDG targets the most well-known online platforms with more than 2 million registered users in Germany. The French laws on information manipulation apply to online platforms whose activity exceeds 5 million unique visitors per month on French territory. The UK Online Harms White Paper suggests a proportionate approach;
- regarding the **type of illegal content online**, the German NetzDG and the French Avia Law are based on national criminal laws. The French laws on information manipulation cover

[99] The NetzDG is available at: https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html.

[100] French draft law to fight hate content on the Internet, as adopted in final reading by the National Assembly, adopted text n° 419, 13 May 2020. The law will enter into force on 1 July 2020. However, the text of the law has not yet been published in the Official Journal of the French Republic. See also Pierrat & Ullern (2019).

[101] Online Harms White Paper, available at: https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper.

deliberate, artificial or automated manifestly false information that could manifestly alter the truthfulness of the election. By contrast, the UK Online Harms White Paper has a very broad scope and covers both illegal and harmful content;

- regarding the **obligations imposed on platforms**, the different laws (or proposals) impose the establishment of a reporting system for illegal content, of internal appeal mechanisms and transparency obligations; in particular for content removal, the French laws on information manipulation impose the removal of terrorist content and child sexual abuse material within one hour; the German NetzDG and the French Avia Law impose obligations to remove within 24 hours the obviously/manifestly illegal content and the NetzDG also imposes to remove other illegal content within seven days; and
- regarding the **enforcement**, the German and French laws provide for sanctions (fines, in particular) for non-compliance; the French laws and the UK White Paper reinforce the role of the regulator.

Table 3: Comparing national laws or initiatives in Germany, France and the UK

Country & Instruments	Germany	France		UK
	NetzDG	Avia Law	Laws on information manipulation	Online Harms White Paper
Category of online platform	Social networks with more than 2 million registered users in Germany.	Online platforms linking several parties for sharing public content and search engines whose activity in France exceeds a certain threshold.	Online platforms whose activity exceeds 5 million unique visitors per month on French territory. Also possibility to impose measures to hosting platforms and Internet providers.	Companies providing social media or online messaging services used in the UK: social media platforms, cloud hosting providers, file hosting sites, public discussion forums, retailers allowing reviewing of products, messaging services and search engines.
Type of illegal content online	22 offences of the German Criminal Code, such as child sexual abuse material, terrorism and serious act of violence, incitement to hatred and violence, hate speech, defamation, some kind of disinformation.	Some offences already existing in French legislation, such as child sexual abuse material, terrorism, incitement to violence, hate speech with a broad definition, sexual harassment.	Deliberate, artificial or automated manifestly false information during election periods that could manifestly alter the truthfulness of the results.	Some online harms, such as child sexual abuse and exploitation, terrorism, revenge porn, hate crime, cyberbullying, disinformation.
Obligations	- Establishment of an effective procedure for reporting illegal content. - Storage of removed content as evidence for ten weeks. - Transparency obligations. - Designation of a representative in Germany.	- Establishment of a uniform reporting system for illegal contents for all online platforms. - Storage of removed content for cooperation with judicial authorities. - Establishment of internal contestation mechanisms. - Transparency obligations. - Designation of a representative in France.	- Transparency obligations during election periods. - Possibility for the judge hearing the summary proceedings to impose measures on hosting platforms and on Internet providers during election periods to stop false information's dissemination. - Put in place measures to combat the dissemination of false information, including the establishment of a mechanism to report false information). - Publication of aggregated statistics on algorithms' operation. - Designation of a representative in France.	Statutory duty of care for companies: - relevant Terms of Service/Terms of Use; - transparent decision-making over actions taken in response to reports of harms; - effective internal complaint mechanisms.

Country & Instruments	Germany	France		UK
	NetzDG	Avia Law	Laws on information manipulation	Online Harms White Paper
Reaction time	- Removal of "obviously illegal content" within 24 hours. - Removal of illegal content within seven days.	- Removal of terrorist content and child sexual abuse material within one hour. - Removal of "manifestly illegal content" within 24 hours.		
Role of regulator		High Audio-Visual Council ensures that online platforms and search engines respect their obligations.	High Audio-Visual Council contributes to the fight against disinformation and ensures that online platforms respect their obligations.	Regulator (Ofcom) should assess compliance with duty of care.
Sanctions	Fines from EUR 500,000 to EUR 5M.	Fines up to EUR 250,000 in case of non-removal of illegal content online. Fines up to EUR 20 million/4% of the worldwide annual turnover for online platforms and search engines that do not comply with measures imposed by the High Audio-Visual Council. Other heavy penalties for Internet users who voluntarily make false reporting.	Fines up to EUR 75,000 for transparency obligations.	

Source: Authors' own elaboration

3. ONLINE MODERATION PRACTICES AND THEIR EFFECTIVENESS

KEY FINDINGS

Online platforms, big and small, **rely on Terms of Service/Terms of Use or Community Standards/Guidelines to regulate and user behaviour and base their illegal content online moderation practices**. These Terms and Standards/Guidelines do not necessarily reflect a specific legal system. However, as they are designed to prevent harm, online platforms' policies do overlap in several instances with local law. These private Codes of Conduct implemented by online platforms may vary from one country to another; they are often **stricter in identifying illegal content online to be removed than national laws** or jurisdictions within which they provide their services.

The **main tools used by online platforms to identify illegal content online are 'notice-and-takedown'/flagging by users, keywords/filters and AI tools based machine learning models**. Most platforms noted that depending on the type of illegal content online, automated tools have their limits in terms of accuracy, and thus, frequently must be accompanied by pre-/post-human moderation to ensure accuracy. The majority of online platforms have argued that the policies put in place by them to moderate illegal content online contribute to **reducing the aggressive nature and the quantity of illegal content online**.

All online platforms interviewed have implemented **transparency policies** on how they operate and respect fundamental rights. Moreover, all of them have **complaint mechanisms** in place for their users to report on illegal content online. However, some platforms have emphasised that many user complaints are off-topic or unsubstantiated and consequently unactionable. Almost all online platforms interviewed allow users to **appeal** against their decisions on the moderation of illegal content online through a 'counter-notice' procedure.

However, most of the interviewed NGOs, trade/industry associations and hotlines reporting illegal content online stated that the measures used by online platforms **are not sufficiently effective in moderating illegal content online and in striking an appropriate balance with fundamental human rights**. Most NGOs and hotlines reporting illegal content online have argued that the effectiveness of the measures deployed by platforms to enable users to report illegal content fluctuates according to the online platform. Additionally, they noted that access to 'notice-and-takedown' procedures is not always user-friendly, whereas they should be easily accessible and not hidden in obscurity.

The **main challenges in moderating illegal content online are linked to the large quantity of online content** on platforms, which makes it difficult for users, regulators or moderators to assess all **content as well as the fragmentation of laws regarding online content**. The Member States are free to set their own rules regarding illegal content online, which limits the efficiency of platforms that have to create country-specific processes accordingly. The lack of a common definition of "illegal content" also makes the moderation by platforms more complex as Member States may refer to different definitions. Therefore, some online platforms mentioned that this places the burden on them to identify the intent of the content uploader, which might incentivise online platforms to block lawful content in case of doubt on the illegality of the content.

Several stakeholders also note that the **current legislative framework on content moderation focuses mainly on the responsibility of online platforms**, while they argue that this should be balanced with rights and obligations of other stakeholders. Most NGOs and industry/trade associations interviewed disagreed with the idea of specific duty of care regimes. They pointed out that **new statutory obligations to remove illegal content online should apply horizontally to any type of illegal content** to avoid regulatory fragmentation.

Almost all online platforms interviewed considered the terms of the **existing liability principles of intermediary service providers** of the e-Commerce Directive as **fit-for-purpose**. However, two indicated that the concept of 'active' and 'passive' intermediary service providers should be reformed. This is because the concept may no longer adequately reflect the economic, social, and technical reality of current services across their lifecycle. Most platforms mentioned that the limitation of liability for Internet intermediaries is a good solution. This is because it allows for protection of fundamental rights, the rule of law and the open Internet.

Regarding **the solutions** to improve the moderation of illegal content online by platforms, stakeholders suggested to put in place **harmonised and transparent 'notice-and-action' processes**. Some stakeholders suggested **strengthening the networks of fact-checkers and hotlines** across the EU. **In terms of fundamental rights, several stakeholders recommended to enforce existing EU rules** and to make them more consistently interpreted across Member States.

Almost all NGOs, industry/trade associations and online platforms interviewed consider that the **existence of different content moderation practices in EU Member States hinder the fight against illegal content online**. Several online platforms stressed that a harmonised approach would enable service providers to have more clarity on what they must do to fulfil their legal responsibilities, while upholding fundamental rights.

In the context of **safeguarding fundamental rights**, most NGOs noted that online platforms' moderating practices should increase moderation transparency, access to data, and information regarding platforms' decision-making processes. In addition, they should ensure human review of the decisions on the user-generated content and contextual expertise. Some online platforms have acknowledged that platforms' incentive to over-remove legal content constitutes the most considerable **threat of unjustified interference** with fundamental rights.

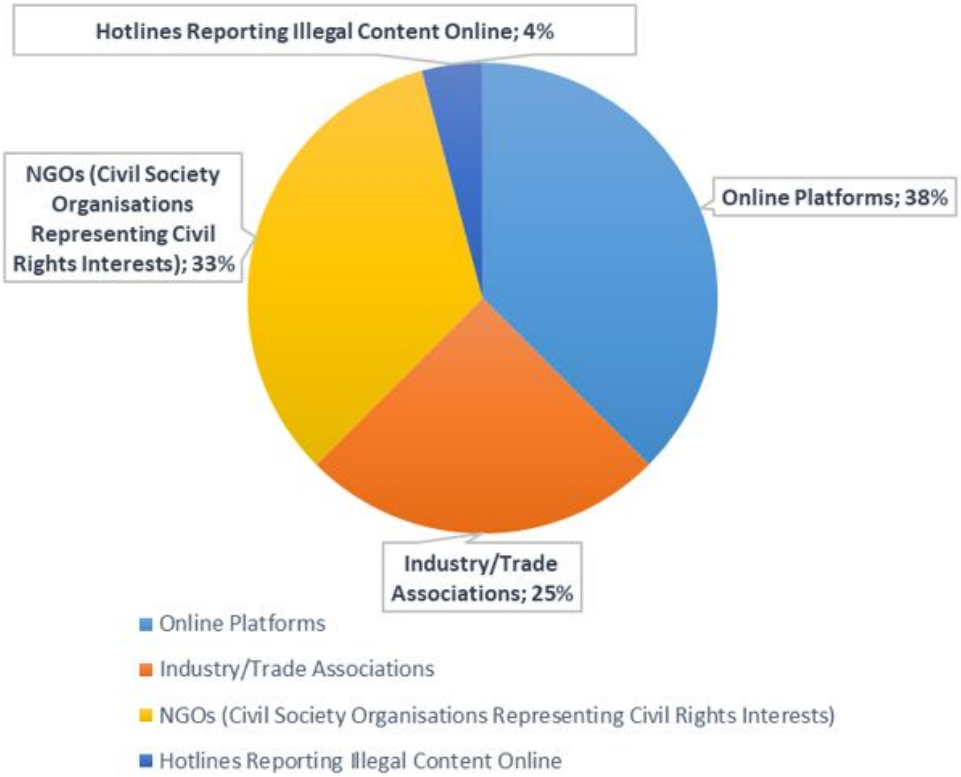
The objective of the stakeholder consultation was to identify the main areas where reforms are necessary due to existing or upcoming barriers or inefficiency/ineffectiveness of current legal solutions in addressing market or regulatory failures.

The first step consisted in mapping the relevant stakeholders to gain information from a broad range of entities.

As a second step, a list of 58 stakeholders was prepared including, online platforms, trade/industry associations, academics, non-governmental organisations (NGOs), and hotlines reporting illegal content online. In total, 24 stakeholders agreed to participate to the consultation and provided their input. This means that for the purpose of this study, nine online platforms, six industry/trade associations, one hotline, and eight NGOs have been interviewed. The following online platforms were interviewed: eBay, Facebook, Google, JustPaste.it, Microsoft, Mozilla, Olx, Snap and YouTube. A complete list of the interviewed stakeholders can be found in Annex III.

As a third step, the study team prepared and sent out two tailored questionnaires: one to online platforms and one to other stakeholders, i.e. trade/industry associations, NGOs, and hotlines reporting illegal content online. A large majority of the interviewees sent written replies to the questionnaire. The graph below presents the percentage of replies to the questionnaire, per type of stakeholders.

Figure 2: Survey replies per type of stakeholders



Source: Authors' own elaboration

During the final step, the answers from the stakeholders were summarised and cross-analysed to draw conclusions on general trends and on the evaluation of the current practices deployed by online platforms and other hosting services providers with regard to illegal online content moderation.

The questionnaires included a set of questions divided into four main thematic groups:

- **measures to moderate illegal content online and their effectiveness**: the objective of the questions listed in this group was two-fold: identifying the measures deployed by online platforms to tackle the illegal content online, and assessing their effectiveness. More precisely, the questions focused on measures aimed at distinguishing legal from illegal content online and at detecting and removing illegal content online. Moreover, the questions enabled to evaluate the impact of these measures and whether there were safeguards in place to ensure that decisions to remove illegal content online are accurate and well-founded;
- **involvement of platforms' users in reporting illegal content online**: the objective of questions listed in this group was to describe complaint mechanisms implemented by online platforms for their users to report on illegal content online. The scope of these questions referred to a 'counter-notice' procedure for content providers; a stay-down issue, namely when the platform should not only remove identified copies of illegal content online, but also detect and remove the same or similar material elsewhere on its platform; and to whether the users of

online platforms are notified when a content they flagged as illegal has been taken down or censored;

- **challenges in moderating/reporting illegal content online and enforcing legal rules, and potential solutions**: the main aim of the questions in this group was to identify areas of the EU Internal Market and its regulatory framework, which require reforms to address existing or upcoming barriers or inefficiency/ineffectiveness of current legal solutions regarding online platforms' illegal content moderation. In addition, these questions concerned the challenges faced by online platforms to enforce legal rules and/or private regimes on illegal content moderation and aimed to identify potential solutions at the EU level to improve the moderation by online platforms of illegal content online; and
- **other issues**: the main aim of the questions in the miscellaneous group was to assess whether the existing liability principles of intermediary service providers (on which Section IV of the ECD is based) are fit-for-purpose, as well as to describe the safeguards mechanisms which would ensure that fundamental rights such as the freedom of expression and information, and the right not to be discriminated, are not infringed in the context of moderation of illegal content online.

3.1. Measures taken by stakeholders and their effectiveness

3.1.1. Moderating measures deployed by online platforms

According to all stakeholders interviewed, **online platforms (regardless if they are a major platform or a small entity) use Terms of Service/Terms of Use or Community Standards/Guidelines** to regulate content moderation activities and the behaviours of its users. As stated by one online platform interviewed, due to their global nature and the extremely broad and diverse community to whom they provide services, their Community Standards/Guidelines do not necessarily reflect any specific legal system. However, as they are designed to prevent harm, online platforms' policies do overlap in several instances with local law. These policies are grounded on feedback from the community and the advice of experts in fields such as technology, public safety and human rights. Importantly, one of the online platforms has indicated that their policies concerning moderation of illegal content online are based on national laws, although in some cases, they may also base such policies on input from public authorities, users and their own discretion, in particular, when unsafe or sensitive items are at stake. The stakeholder in question has, in the light of consumer protection, committed to additional voluntary requirements to ensure the safety of products that are sold via that platform by signing the EU Product Safety Pledge. This initiative provides for dedicated voluntary actions which go beyond what is already required by EU legislation.

The private Codes of Conduct implemented by online platforms may vary from one country to another and **are stricter in nature in identifying illegal content online to be removed than national laws** of a relevant jurisdiction within which they provide their services. According to some NGOs interviewed, online platforms tend to remove more content than just what is illegal under the law of a given jurisdiction within which they operate, to stay on the safe side, as they are not equipped with the same tools which are at the disposal of the Courts. In addition, their knowledge and incentives to carry out a proper legality assessment and balance different rights at stake are limited. This means that online content removals are mostly based on online platforms' stricter Terms of Service/Terms of Use rather than the legal rules of a given country.

In terms of specific measures that have been deployed by online platforms to distinguish legal from illegal content online, the **main tools used to identify illegal content** are 'notice-and-

action'[102]/flagging by users, machine learning models to replicate human decision making, keyword filters and the removal of entire content from a given platform.

Regarding the **'notice-and-takedown'/flagging mechanism**, several online platforms indicated that this tool is deployed as it is not possible to detect the majority of illegal content online by technical means and is impractical to do so at scale through human moderation. As contended by one of the platforms, although the 'notice-and-takedown' process is the same, the set of information which needs to be provided by notice providers to the platform's dedicated team differs according to the various types of infringements. As a result, the actions taken in response to illegal content also differ according to the type of illegal content. This is the case both in terms of speed of action against specific types of content and of the degree of severity of measures deployed which generally go beyond local law.

In terms of automated content moderation, one of the online platforms highlighted that automated tools could "*damage user experience by over-detection and the generation of false-positives*". Consequently, automated content moderation cannot be treated as 'a complete solution', as most content moderation requires human verification, leading to significant economic and human investments on the platforms' side.

In order to ensure that decisions to remove illegal content from a platform are accurate and well-founded, especially when automated tools are used, **online platforms use human moderators** and treat their decisions as superior to those made independently by machines. One of the platforms emphasised that the main advantage of human moderators is that their review will always allow for a greater degree of context and common sense to be applied to the online content in question. Elements of context such as local culture, traditions, politics, etc., play an important role for edge cases where it may not be fully clear whether the platform's Terms of Service/Terms of Use or Community Standards/Guidelines have been breached.

Regarding the **speed of action to take down illegal content**, the majority of platforms interviewed usually remove terrorist-related content and material of child sexual exploitation within one hour, while any other illegal content online is removed within 24 hours.

All online platforms interviewed have stated that they have implemented **transparency policies** to ensure full transparency in how they operate and respect fundamental human rights. Consequently, online platforms have certain standards in place for how they communicate with users when informing them about moderation practices which have been taken in relation to content they have posted. There are also standards for how platforms communicate with third parties who notify allegedly illegal content online to them. Moreover, online platforms publish on a regular basis transparency reports which provide insights into volumes and types of requests regarding online content moderation they have received, for instance, from law enforcement authorities. Furthermore, one of the online platforms has recommended not to disclose algorithms, rules-based filters or other factors, that would "*inevitably allow fraudsters to game the protective systems*" deployed by online platforms.

3.1.2. Online platforms' perspective on the effectiveness of the deployed moderating measures

With regard to the automated moderation of content that appears online, six out of nine online platforms indicated that the **use of automated tools for moderation of illegal content online brings efficiency** into practice and is valuable in dealing with large volumes of data. The artificial intelligence-

[102] 'Notice-and-action' is an umbrella term for a range of mechanisms designed to eliminate illegal or infringing content from the Internet. It comprises mechanisms such as the 'notice-and-takedown' scheme which currently results from Article 14 of ECD.

based tools used by platforms include upload keyword filters, machine learning models and shared industry databases of hashes (or 'digital fingerprints') to increase the volume of content which can be caught at the moment of uploading by their machines.

However, most platforms also pointed out that **depending on the type of content**, **automated tools have their limits in terms of accuracy**, **and thus frequently must be accompanied by pre-/post-human moderation** to ensure precision. One of the online platforms acknowledged that, for instance, in the context of counterfeit goods, even experienced and highly specialised human moderators do not always have the expertise to make 100% accurate decisions because of the high number of products, manufactured by an enormous number of brands, for them to be able to be precise in their analyses. Thus, such online platform has to rely heavily on the external expertise of specialised NGOs, brand owners, or other competent third parties. Seven out of nine online platforms interviewed use automated moderating tools based on artificial intelligence (AI) and most of them strive for human review of the content that has been removed.

The majority of online platforms argued that the policies put in place by them to moderate content contribute to **reducing the aggressiveness and quantity of illegal content online**. According to one of the platforms, the moderating policies and the tools used are "*certainly successful*" in stopping large quantities of illegal content from appearing online. However, at the same time, this stakeholder pointed out that criminals are innovative and are continually coming up with new ways to circumvent the oversight measures deployed by platforms. As a result, online platforms are in a constant 'arms race' against cyber criminals. Regarding the removal of videos containing illegal content, in 2019, one of the online platforms interviewed removed more than 30 million videos for violating their Community Standards/Guidelines. Furthermore, another platform has noted that in 2018 more than 2.3 billion 'bad ads' were removed from their systems due to violating their advertising policy. Furthermore, in 2018, they intervened against almost 1 million bad advertiser accounts. Additionally, with regards to unsafe products, one of the platforms has disallowed 5 million listings that were identified via monitoring of public recall websites, namely, RAPEX, for the period of April-September 2019 under the Product Safety Pledge.

Importantly, it has been noted that successful and effective work against illegal content is **impossible for online platforms to achieve in isolation**. Thus, the success equally depends on the cooperation and the sharing of data and knowledge with other key digital market players, such as Internet providers, third parties and national law enforcement authorities.

3.1.3. Other stakeholders' perspective on the effectiveness of the moderating practices

From the perspective of most of the NGOs, trade/industry associations and hotlines reporting illegal content online, the measures used by online platforms **are not sufficiently effective in moderating illegal content online and striking the balance with fundamental human rights**.

One of the NGOs pointed out that not all online platforms are at similar levels in terms of quality moderation because they have **very different sets of internal policies applied to moderate illegal content online**. It was stated that there are more and more similar initiatives developed by well-established platforms such as Facebook, Twitter or YouTube, while others such as TikTok, Yobo or Snap, for instance, use different protocols, which differ in terms of AI investments or partnership with NGOs.

One of the NGOs also mentioned that the fact that online platforms apply their own standards makes them "*biased by definition*". It has been emphasised by most of the interviewed platforms that there is no one-size-fits-all approach in relation to content moderation measures. This is because various online

service providers play fundamentally different roles and have divergent responsibilities. Online platforms provide a variety of products and services, and the measures they take may vary accordingly.

In addition, according to one of the NGOs interviewed, many of the largest platforms, like Facebook, Twitter, YouTube, TikTok, **do not involve enough their end-users or civil society**, beyond trusted flaggers. Facebook and TikTok are currently assembling advisory councils, but it remains to be seen whether these councils will meaningfully involve end-users. Wikimedia was mentioned as an example of a larger platform, which does implicate users in the process of moderating content, and it does so through multiple mechanisms that encourage users to increase their moderation responsibilities and become members of the community[103].

Furthermore, two other NGOs estimated that the **transparency reports published by some online platforms are not complete enough** to assess the effectiveness of the measures deployed to moderate illegal content online. The data published by larger platforms (such as Facebook, YouTube, Twitter, Wikimedia, LinkedIn, Reddit, Jeuxvideo.com) in their transparency reports demonstrate that they have been effective at responding to illegal content online. However, the NGOs at stake argued that this data does not provide them with clear information on what type of content is taken down and on what basis. One of the NGOs added that they lack information regarding false-positives, and regarding the migration of illegal content online between platforms, i.e. "*a platform may be successful at eschewing illegal content, but what damage does that content cause elsewhere?*". Importantly, most of the interviewed NGOs highlighted the "*porosity of illegal and legal content*" which may lead to incorrect classifications of content online or the failure to identify content (false-positives or false-negatives). Several NGOs also mentioned that there are a number of cases where legitimate content online was blocked by social media platforms, while, on the other hand, reports concerning clearly harmful and illegal content online have been ignored.

3.2. Involvement of platforms' users in reporting illegal content online

3.2.1. Online platforms' perspective

All online platforms interviewed have complaint mechanisms in place for their users to report on illegal content online. **The majority of online platforms interviewed used reporting mechanisms based on flagging tools, while only one uses email alias** as a direct reporting channel for users. The different flagging tools for users consist, depending on the platform, of:

- a dedicated button or link in the listing;
- a simple flagging tool allowing a user to press down on any item of online content (i.e. private or public), following which a flag appears, alongside a series of categories of common content complaints (e.g. nudity or sexual content, hate speech, threatening, violent or concerning, etc). Users can select one of these categories, or just indicate that they do not want to see such content. Users may provide more information via a free text box; or
- a dedicated on-platform support inbox, used to share information about the status of the report.

Moreover, three of the online platforms interviewed indicated that they deploy separate mechanisms for flagging potential violations of their Terms of Service/Terms of Use or Community Standards/Guidelines, and for flagging content, which users believe violates national law and is not otherwise covered by platforms' policies. This two-fold distinction of the mechanisms results from the

[103] Details of this initiative of the Wikimedia Foundation are available at: https://wikimediafoundation.org/participate/.

fact that the online platforms' Terms of Service/Terms of Use or Community Standards/Guidelines are developed for a global user base and are not intended to cover all local laws.

Regarding the first type of mechanisms on infringements of online platforms' Terms of Service/Terms of Use or Community Standards/Guidelines, the user can choose from different content categories to select the reason why they are reporting the content. Moreover, one of the platforms has also developed a 'Trusted Flagger' programme to help encourage simultaneous submissions of multiple high-quality flags about content that potentially violates their Community Standards/Guidelines. The role of trusted flaggers is played by NGOs, government agencies and individuals who have a high accuracy rate and domain expertise that makes their flagging a valuable input for the overall system. According to one of the platforms, the majority of users' reports are reviewed within 24 hours. To achieve this, the platform uses a combination of human review and automation. If the reported online content violates their Community Standards/Guidelines, the platform takes it down, otherwise, they leave it up. If the online platforms at stake delete users' content for violating Community Standards/Guidelines, they inform the posting users of their action and that the content violated the Community Standards/Guidelines.

With regard to the second type of reporting mechanisms, i.e. concerning legal complaints, the same three online platforms indicated that the users who intend to complain have to complete the appropriate form, which helps to ensure that the platforms have all necessary information to investigate a specific enquiry and resolve it as quickly as possible. One of the online platforms interviewed has provided a list of legal reporting channels available to platform's users, which includes:

- publicly accessible Intellectual Property reporting channels allowing Intellectual Property Rights-holders to report online content they believe violates their rights, including copyright infringement, trademark infringement and counterfeits. These channels feature dedicated reporting forms for each type of infringement, as well as a dedicated email alias for reporting;
- a defamation reporting form allowing users to report content they believe is defamatory under local law;
- a legal removal request form allowing individuals in EU Member States to report content they believe violates local laws; and
- a reporting form allowing users in Germany to report content they believe violates one or more of the German Criminal Code provisions.

One of the platforms has stated that allowing users to easily access an intuitive reporting flow adjacent to the content may lead to a **high number of clicks and complaints that are often unreliable**. This is due to the fact that many of the users' complaints are off-topic or unsubstantiated and consequently unactionable. There are also users who submit a complaint without providing any information on why they think the content is illegal. In addition, one of the platforms at stake has noted that according to their analyses of cease-and-desist and takedown letters, many users seek to remove potentially legitimate or protected speech. Importantly, this stakeholder has also pointed out that a vast majority of complaints, regarding alleged infringements of the German Criminal Code, are unsubstantiated, regardless of the fact that the platform has explicitly asked for further details[104]. Therefore, if the analysed online content does not violate the platform's Community Standards/Guidelines, the platform will conduct a careful legal review to confirm the validity of the report. In cases where reports

[104] More precisely, the platform in question has noted that 74% of content reported under the form dedicated to users in Germany were not removed or blocked, since the content has neither constituted an infringement of their Community Standards/Guidelines nor the German criminal law.

are not legally valid, overly broad, or are inconsistent with international norms, a request for clarifications will be sent or no action will be taken by the platform. If the content is found to indeed violate local law, the platform will make it unavailable in the relevant country or territory and will publish the aggregate information about the requests in their transparency report.

Another online platform, acting as an online marketplace, has stated that they treat the "*vast majority of third-party notices (e.g. from brand owners about counterfeit goods or from wildlife NGOs about CITES-protected animals or animal products) as justified*". This is because their expertise is limited to be able to dispute the claims. According to a rough estimate of that platform, the number of notices of allegedly illegal content that they reject is in the low single digits in terms of percentage of total notices received. In this context, the platform in question has added that "*while this percentage might appear negligible, each unjustified action against user content creates the potential for lengthy, costly, and reputation-damaging disputes*".

Eight of the interviewed online platforms allow users to appeal against their decisions on content moderation through **a 'counter-notice' procedure**. However, two of those platforms do not notify the users when the content they have flagged as illegal has been taken down or censored, but only display a message on a content page stating that the content that has been removed. The platforms interviewed indicated a number of advantages and disadvantages of having content providers being able to give their views to platforms on the alleged illegality of the content through a 'counter-notice' procedure. The advantages of 'counter-notice' systems include:

- greater content provider satisfaction because of a perception that their interests are protected;
- a decreased likelihood of spurious or unjustified claims by users flagging the content online;
- a higher likelihood of accurate decisions concerning content moderation due to appeal requests which are reviewed by human senior reviewers who did not make the original decision to remove the content in question; and
- the removal of illegal content that is linked to one platform but hosted on the other.

The disadvantages are associated with online platforms becoming akin to dispute resolution bodies and include:

- the legal risks associated with becoming a party to a dispute between a user and a third party;
- the resources required to mediate (i.e. time, systems, tools and external legal advice); and
- the risk for the protection of identity and anonymity of users who flagged illegal content during the course of a 'counter-notice' procedure.

As it has been emphasised by one of the platforms, such risks only increase when the user has flagged content posted by violent individuals or groups.

Among the nine online platforms interviewed, only three provided information on how long the measures they take against illegal content online remain effective and the way the **stay-down issue** is addressed by them. Generally, platforms either prevent upload of individual pieces of content permanently, or remove the content permanently if it is caught after upload. Their moderating systems are able to identify duplicated (i.e. identical) content, for instance, if a user tries to re-upload data that has already been removed. This is usually done by using hashes to catch copies of known content before they become accessible to Internet users or by preventing exact match URLs from being re-indexed. However, one of these platforms has clearly stated that from a technical perspective, it is very complicated for them to block similar online content, in particular at scale. Consequently, they argued

that in practice it is very problematic to apply stay-down measures requiring the removal of content, which is similar, but not an exact match. Therefore, this regime would be impractical and cause "*a massive wave of litigation*", because of inadvertent blocking of legal content and the risk of overly broad removals. The current regime strikes a good balance in this field, as can be inferred from the views of these three online platforms.

3.2.2. Other stakeholders' perspective

The vast majority of other stakeholders interviewed, namely NGOs, industry/trade associations and a hotline indicated that **online content providers should be able to give their views to the hosting service on the alleged illegality of the content through a 'counter-notice' procedure**. Most of the NGOs interviewed stressed that a 'counter-notice' procedure constitutes one of the crucial mechanisms to prevent content over-blocking and that 'counter-notices' should be part of a larger users' recourse process that is transparent, accessible and timely. One of those organisations stated that the procedure in question should be used in cases involving complex legal assessment, such as Intellectual Property-related issues. This allows to have greater safeguards against abusive notices and avoid suppression of legitimate content. Furthermore, according to another NGO interviewed, a 'counter-notice' procedure serves as a significant tool for protecting freedom of expression and diversity of opinion online, as well as fair competition in the Digital Single Market. This stakeholder also pointed out that it should be a basic right of content providers to defend their viewpoints and interests in dispute settlements, unless inappropriate, and to contest the opinion of the notifier.

Most of the interviewed NGOs and hotlines reporting illegal content online argued that the **effectiveness of the measures deployed by platforms to enable users to report on illegal content is fluctuating depending on the online platform**. For example, according to one of the hotlines dedicated to Intellectual Property Rights, the reaction time and the intervention of Facebook varies significantly when they have reported on closed user groups with illegal services and products. The same experience applies to Google when calling for the removal or demotion of illegal search results. In relation to the share of content taken down as a result of the reporting by users, five out of nine online platforms indicated that the percentage of the illegal content taken down is low and most of the content that is removed is first detected by automated flagging mechanisms or by human moderators. More precisely, one of these five online platforms noted that between October and December 2019, they removed over 5.8 million videos for violating their Community Standards/Guidelines. Machines rather than humans first flagged 90% of these removed videos. Of those detected by machines, 64.7% had never received a single view. Considering illegal content online taken down by a platform due to the users' notifications/complaints in comparison to other sources of first detection, another interviewed platform added that the proportion at stake is "*probably also in the low single digits, or even less than 1%, of the total number of notices we receive*". Apart from the aforementioned online platforms, the other online platform noted that between July and September 2019, 98.4% of illegal content online actioned was found and flagged by the platform before users reported it. Thus, the percentage of violating content that users reported first between July and September 2019 was only 1.6%.

Many stakeholders have also noted that accessibility of the **'notice-and-takedown' procedures is not always user-friendly**, while such mechanisms should be easily accessible by users and not hidden in obscurity. One of the NGOs, concerning major social media platforms, pointed out that users **have little possibility to contest** online platforms' decisions on content moderation, and they fail to provide to users "*a due process and effective remedies for wrongful removal*" of online content. In the context of a

due process, one of the interviewed organisations highlighted that EU Member States should provide "*an avenue of appeal*" once the internal mechanisms of social media companies have been exhausted.

The aforementioned NGO also observed that platforms can always "*hide behind their unclear Terms of Service*" to justify removing any content, referring to the Council of Europe 2016 study[105] in this context. The study concludes that "*(…) the majority of platforms (52%) explicitly state that they may remove content based on third-party notification without offering any justification, notification or opportunity to be heard to the user who originally shared it*" and "*(…) there is also little commitment to offering users justification, notice and the right to be heard when content is removed by the platforms' own initiative or after notification from third parties*"[106].

Regarding the type of measures which should be taken to **improve the transparency of platforms' decisions** concerning illegal content online reported by users, only one stakeholder (i.e. an industry association) stated that the current framework is "*appropriate to a high level of transparency*" since many of its members already publish reports on their content moderation practices and outcomes. For example, the aforementioned stakeholder indicated that Amazon releases transparency reports on law enforcement information requests, while Facebook publishes *Community Standards Enforcement* reports. Moreover, it was also added by that stakeholder that Google shares transparency reports on content removal that comprise data on content delisting due to copyright or to the enforcement of YouTube's Community Standards/Guidelines. Importantly, this group of stakeholders, namely NGOs, hotlines reporting illegal content online and other trade/industry associations than the aforementioned one, argued that there is not one single best practice for transparency reporting. At the same time, one of the NGOs indicated that imposing overly strict transparency reporting requirements on platforms of different capacity and nature would not give useful results. The vast majority of the aforementioned stakeholders have indicated that:

- online platforms should implement transparent (i.e. clear, accessible, understandable and specific) Terms of Service/Terms of Use or Community Standards/Guidelines, which should be available in all languages in which the services are offered;
- online platforms should inform users when a moderation decision is made on their content and they should include adequate information on what triggered the decision, the specific rule that has been infringed, how the content moderation guidelines were interpreted, the actions that will be taken, and clear instructions for an appeal;
- users should have a possibility to effectively appeal from the platform's decision. The recourse process should be easily understandable and accessible; and
- online platforms should be obliged to regularly publish transparency reports. These reports should at a minimum contain comprehensive information about notices, the types of content to which they relate, the notifiers, appeals, and staff employed for content moderation.

Other suggestions on how to improve the transparency of online platforms' decisions concerning illegal content reported by users include, according to one of the NGOs: (i) "*a stronger Moderation Policy and a Charter*" which would serve to explain to the users what is considered or could be considered as illegal content, including relevant examples; and (ii) a common button and procedure for all platforms. In the stakeholder's opinion, this would allow users to know only one procedure to signal and report

[105] Venturini and al. (2016).
[106] Venturini and al. (2016).

illegal content, as well as to allow users to follow common guidelines, thus ensuring that illegal content is understood equally by everyone.

3.3. Challenges in moderating illegal content online

3.3.1. Challenges in moderating and reporting illegal content online and enforcing legal rules

An important challenge in reporting illegal content online, mentioned by a few of the NGOs interviewed, is the **large quantity of online content on platforms**. They stressed that there is too much content for users to report on and that regulators are hampered by constraints with funding, resources and adaptability. These NGOs indicated that 'notice-and-takedown' procedures, requiring monitoring and reporting, should thus not be left to companies, users and regulators.

Another challenge in reporting online content lies in the fact that **average users are not necessarily capable of accurately identifying illegal content online**. An NGO mentioned that most users "*are not specialised lawyers and therefore the challenge for them is how to adequately identify a piece of content as illegal*". In practice, this would mean that average users can usually only make an assumption or express a suspicion about the potential illegality of content online.

According to the same NGO, another challenge is linked to the **accessibility of reporting tools**. The NGO claimed that online platforms discourage people from using them by making explicitly inhibiting design choices for the user interface (so-called 'dark patterns' that manipulate the user to behave in the platform's interest). According to the NGO, an evaluation of the implementation of the German Network Enforcement Law showed that the complaint form was relatively hard to access, on Facebook in particular [107]. It was added, for example, that if all notices/reports require to identify oneself with their real identity (national ID number, name, address, etc.), even when it is not necessary to process the notice, there is a "*potential chilling effect to the use of such reporting system*", especially for cases involving hate speech (fear for retaliation, discrimination, etc.).

The **online platforms interviewed noted that they are continuously facing new trends and challenges** in online content moderation. One of the main challenges they face in enforcing legal rules on moderation of illegal content online, as mentioned by several NGOs and online platforms, lies in the fact that they have to deal with **multiple legal frameworks** (in their home country, in their countries of implementation and in harmonisation between Member States of the EU). These stakeholders agreed that having to deal with various prescriptive direction from individual Member State governments complicates a platform's ability to be efficient in how it allocates resources. As reported by one online platform, an example of fragmentation is the difference in scope and application between the German *Network Enforcement Act* (NetzDG) and the French Avia law. For instance, whereas the German NetzDG is limited to hate speech and restricts the application to certain social media services where the risk of proliferation is high, the Avia law includes an expansive definition of hate speech that depends heavily on context. In both cases, these Regulations impose "*significant country-specific reporting and response requirements that obligate the platform to create country-specific processes which limit their efficacy and scalability*".

[107] To defend its argument, the NGO quoted the study, *An Analysis of Germany's NetzDG Law*, published in April 2019, by Heidi Tworek and Paddy Leerssen: "*YouTube and Twitter integrated NetzDG complaints into their regular "flagging" interface, which can be accessed through direct links next to every piece of content. Facebook, by contrast, placed their complaint form on a separate, less prominent page, requiring multiple clicks to access. […] The report data suggest that this design choice had massive impacts on the actual uptake of NetzDG complaints.*" Study available at: https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.

Moreover, and as mentioned by an NGO, managing varying content laws by jurisdiction forces services to apply the most restrictive content policies worldwide. The consequence of this is a jurisdictional conflict resolved only by "*overly inclusive restriction of speech, or by limiting access to services by geography*", which undermines "*free and fair access to the open Internet*".

The lack or vagueness of a definition of illegal content was also mentioned, by one online platform and several NGOs interviewed, as issues in moderating illegal content online. Vagueness in definitions of certain content, such as hate speech or extremism, place a burden on technology companies to attempt to ascertain the intent of the speaker often with little context. The definitions of such targeted content are not even commonly agreed among researchers, who are using varying definitions in their studies. Furthermore, it was mentioned that slang, abbreviations, symbols, and other imagery can have "*many legitimate uses absent context or clarity of intention – neither of which may be available to or understood by artificial intelligence or human moderators*". In addition, it was stressed that when the penalties for failure to remove hate speech are significant, technology companies are "*forced to err on the side of blocking lawful content*".

An online platform and several industry/trade associations pointed out that the main challenge that platforms have in moderating illegal content online is the fact that the **current legislative framework is focused on the responsibilities of platforms and "*does not provide enough structure to the rights and responsibilities of others*"**. The online platform stated that, for example, public authorities do not always make available databases of important (and theoretically public) information that would help them comply with the law. The online platform added that they have occasionally felt that some brand owners (ab)use the 'notice-and-action' (N&A) system to enforce distribution agreements. This latter concern was shared also by an industry association.

Another challenge mentioned by a few of the NGOs interviewed is that online platforms are **asked to act expeditiously to remove illegal content but to not interfere with the transmission** (i.e. not to be involved in the processes which would give effective knowledge of the potential illegality of the content) in order to benefit from safe harbour protections provided by the e-Commerce Directive. According to these NGOs, this may create "*a strong incentive for online platforms to either 'look the other way' or to over-remove everything that could even remotely be illegal under any jurisdiction somewhere in order to avoid legal liability*". Consequently, according to these NGOs, it is highly problematic to leave it to the private sector to decide over the proper balance between fundamental rights, as this leads to arbitrary decisions, in particular when the incentives are imbalanced (risk of liability, reputation, authority of the notifying party, etc.). The NGOs at stake also questioned whether intermediaries can reasonably be asked to make such rulings of (il)legality based on assumptions/notices made by third parties (i.e. other users) before the content provider has had the chance to defend themselves.

Finally, the last key challenges mentioned in this context by an online platform are the **Europe's data protection rules**, in particular the e-Privacy Directive. It was mentioned that, when scanning for illegal content online, particularly child sexual exploitation and abuse, online platforms "*need to balance users' privacy rights with child safety and law enforcement/government requests particularly, in geographies with strict privacy and communications secrecy laws where a breach may result in financial penalties or criminal charges*". The challenge lies in the fact that the European Union and some Member State laws[108] prohibit interference with online private communications, rendering scanning for child sexual exploitation and abuse material on online platforms, legally questionable.

[108] Germany, the Netherlands, France, Spain, and Italy have similar communications' secrecy laws, stemming from both constitutional rights to privacy of communication and the e-Privacy Directive (2002/58/EC) that require service providers to obtain explicit consent by one or both parties.

3.3.2. Duty of care regimes

A few of the NGOs interviewed agreed that a specific duty of care regime for certain categories of illegal content online is needed. However, it was stated that it should not be accompanied by "*strict requirements and fines that ultimately encourage platforms to remove content more heavily*" than if a duty of care did not exist. According to these NGOs, the duty of care should require platforms to show that "*they are making good faith efforts and to give visibility into those efforts but need not obligate certain results*". It was suggested by one NGO that, in case of failure to perform its duties of moderation, the responsibility of the platform to delete illegal content must also be searched on the ground of its online duty of care that it must guarantee to all users. Therefore, in case of a conviction, the sanction should be aggravated if the Court establishes that the content in question should have required a duty of care beyond doubt.

To increase the transparency of measures and the duty of care regime, an NGO proposed that two main categories of intermediaries have to be established beside the duty of care of the platform:

- common users should be made aware of the duty of care and how it works as well as the promotion of an active moderation toward them when they register to the platform; and
- public or private entities using the platform as a medium for public outreach should be considered as well-advised and prepared users able to carry out any action necessary to the moderation of their contents, posts, comments and sponsored spaces. The duty of care should be ratified by them and should hold them responsible before Courts in case of failure to moderate.

Regarding the question of a duty of care specific for certain types of illegal content, on the one hand, some stakeholders (two NGOs) considered that all categories of illegal content should require a specific duty (classifying content into different categories risks "*creating black boxes without sufficient oversight and which risk false-positives and false-negatives*"). On the other hand, two other NGOs believed that there should be different categories of illegal content with regards to the duty of care. An NGO proposed that hate speech, incitement to violence, and child pornography should specifically require duty of care.

However, most NGOs and industry/trade associations interviewed disagreed with the idea of specific duty of care regimes. One NGO mentioned that, in their experience, duty of care regimes are "*usually poorly defined and end up requiring intermediaries to proactively monitor, judge, and remove potentially illegal user and third-party content on networks and online platforms in order not to lose liability exemptions*". Mostly, they take the form of political pressure on platforms to take (formally voluntary) measures without clear and understandable obligations, and to set up predictable sanctions for failure to comply with these measures. An NGO claimed that this duty of care can be considered contrary to Article 52 of the EU Charter of Fundamental Rights, under which "*restrictions on fundamental rights must have a legal framework which is sufficiently clear to enable both natural and legal persons to regulate their conduct*".

Several industry/trade associations indicated that new statutory obligations to remove illegal content online should apply horizontally to any type of illegal content to avoid regulatory fragmentation. Differing procedures for different types of content should be justified by objective distinctions (for example, whether or not the nature and legal status of the content is objectively classifiable; whether or not the alleged infringement is of criminal law or instead of private rights). According to one of the industry/trade associations, diverse duties of care across different services that are already the subject of many varying definitions (Video-Sharing Platforms (VSP) in the Audio-Visual Media Services Directive(AVMSD), Online Content-Sharing Service Providers in the Copyright in the Digital Single

Market Directive (CDSMD), Internet access and Service Providers under the e-Commerce Directive (ECD), etc.) "*will only add to layers of parallel liability regimes without creating any legal certainty for service providers to actually take action*".

3.3.3. Solutions to improve the moderation of illegal content by online platforms

Some of the online platforms expressed that a **more inclusive legislative framework should be put in place**, in which all stakeholders (online platforms, users, brand owners, law enforcement authorities) have clear and balanced rights, obligations and liabilities. It was also proposed that new responsibilities should be considered for all stakeholders, including for example clarifying the liabilities of notice providers for damages done to users' rights and the interests of platforms.

An **increased use of fact-checkers** was also mentioned by an industry association as a way to improve moderation of illegal content online. There should be some in every country and the source of the content should be made more transparent, and if the source is not reliable, then the content should be limited.

Regarding an increased use of automated tools, several online platforms mentioned that automated tools can "*bring efficiency into the content moderation practices*", but they are not always accurate or reliable, especially depending on the type of content moderated. In line with that, many stakeholders (including online platforms and NGOs) mentioned that a human oversight is inevitable. One online platform was even more critical regarding automated tools, as it stated that they can produce "*serious interference with individuals' fundamental rights and cement the power of the handful of large companies who have the technology and resources to comply with filtering mandates*"; at a minimum, automated tools "*should not be mandated by law according to the platform*". To that end, the same stakeholder recommended that the *Digital Service Act (DSA)* should include requirements or incentives for companies to implement effective due process into their deployment of automated content moderation technologies, along the lines of those suggested in the *Santa Clara Principles*[109].

Regarding new measures to improve the EU regulatory framework and its enforcement regarding online platforms' moderation of illegal content online, an NGO mentioned that the **EU regulatory framework needs indicators that will assess processes**. Indicators should be used to assess platforms' responsiveness to problems like illegal content. This stakeholder suggests that, rather than focusing on the threshold of users or on any specific technical feature, regulation should evaluate platforms on processes, in particular, those that relate to transparency and accountability, dialogue with users, inclusion of civil society and protection of fundamental rights and freedoms.

Several stakeholders, including NGOs and a trade association, agreed that the **'notice-and-action' mechanism should be reformed and harmonised**. One of those NGOs proposed to implement a notice-and-action system, with specific quality criteria for notices including:

- the name and contact details of the notifying party only in cases where this is necessary to process the notice (mainly for copyright and defamation cases);
- the link (URL) or a similar unique identifier to the allegedly illegal content in question;
- the stated reason for the complaint including, where possible, the legal basis the content in question is allegedly infringing;
- depending on the type of content, additional evidence for the claim; and

[109] Santa Clara principles on transparency and accountability in content moderation, available at: https://santaclaraprinciples.org/.

- where a complaint is not anonymous, a declaration of good faith that the information provided is accurate in cases of copyright infringement and defamation cases.

In order to make the notice-and-action system workable, the required online notice forms should be straightforward to use and easily accessible. It should also allow content providers to issue a 'counter-notice' to defend their viewpoint and interests, except where such a 'counter-notice' would conflict with an ongoing criminal investigation, which requires to keep the decision to suspend or remove access to the content a secret. For example, child sexual abuse material should be made inaccessible as quickly as possible, while notices of alleged copyright infringements or defamation need to provide the content provider with sufficient time to react before the content in question is removed.

The same NGO also suggested **implementing an alternative dispute settlement**, with online content dispute tribunals. It stated that, in order to facilitate access to remedies for users in the face of powerful online platforms, "*the EU should require or at least encourage Member States to establish independent dispute settlement bodies for users in their jurisdiction*". These independent bodies should serve as a tribunal system providing simplified legal procedures tailored to the nature of online content moderation disputes. Their task should be to settle disputes between users, as well as with all online platforms, regarding the legality of user-uploaded content and the correct application of Terms of Service/Terms of Use when they relate to content moderation decisions taken by online platforms.

Several NGOs and two online platforms suggested that the measures deployed should be **differentiated according to the type of illegal content and the type of services (size, scale, function) displaying the content**. One of the NGOs suggested that the EU regulatory framework should distinguish between disinformation, hate speech, copyright violation, defamation or libel, or terrorist content. One of the online platforms in question added that the measures should differentiate, for example, between services whose "*primary purpose is to make content widely available to the public by default, and those that are used primarily for personal storage of private content and are not designed to facilitate broad dissemination of content*". One of the NGOs gave the example of a small start-up with minimal user-generated content, which in their opinion, needs different moderation practices than a publication platform aimed at children or extremist groups. Moreover, regulation should consistently differentiate "*between service categories that bad actors frequently use to disseminate terrorist content and services that they rarely rely on for such purposes*".

To improve the EU regulatory framework and its enforcement regarding online platforms' illegal content moderation, an NGO also suggested that the **European framework should strengthen hotlines at national level** so they can "*act as public-private partnerships or at least as a national hub for receiving and assessing reports of illegal content*". In the US, the congressional mandate given to the National Centre for Missing and Exploited Children (NCMEC), combined with the obligation for online platforms to report illegal content to NCMEC, has proved to be an efficient model, as stated by several NGOs. It was noted that there is a valuable set of hotlines in the EU but that the model remains embryonic and limited in scope (not all hotlines cover hate and terrorism, and none cover copyright infringements). As stated by an NGO: "*A robust European legal framework organising and empowering national multi-stakeholder hotlines would facilitate the regulation of illegal content online, to the benefit of the platforms, the regulators, and ultimately the victims of abuse and the citizens*".

Finally, several stakeholders including online platforms and NGOs stressed that the **Digital Services Act is an opportunity for the EU** to adopt a new, more effective regulatory paradigm. The DSA framework should be "*targeted, proportionate, and infused with the right incentives for platforms to address why and how illegal content disseminates through their services*".

Reforms in areas of the EU Internal Market are necessary to address existing or upcoming barriers or inefficiency/ineffectiveness of current legal solutions regarding online platforms' illegal content moderation. Two of the online platforms interviewed mentioned that one of the main areas of the EU Internal Market that should be reformed is the **area of Data Protection**. It was suggested to rely mainly on the General Data Protection Regulation (GDPR) and to reform the e-Privacy rules to make it explicit that "*scanning for the purposes of implementing PhotoDNA or equivalent tools should not require user consent and will not represent a breach of communications secrecy laws*". This would enable to deal with the issue mentioned previously that some Member State laws prohibit interference with private online communications and, hence, may render scanning for CSAM on online platforms, legally doubtful.

Reforms in the **area of the Digital Single Market** were also suggested by a few of the NGOs and online platforms, notably regarding the **'country of origin' principle**[110], which should be strengthened in all Internal Market instruments when they come up for review, with derogations being removed wherever possible (ECD, AVMSD, CDSMD, etc.). An NGO and three online platforms stressed the benefit of the 'country of origin' principle, which is fundamental for the development of the EU Internal Market and the facilitation of cross-border trade. According to one online platform, the principle makes possible the free movement of goods and services within the EU by ensuring that information society services are "*supervised at the source of the activity*". The stakeholders at stake strongly recommend maintaining the principle, without which European companies might face increased challenges to scaling up in the EU Internal Market. For one of the NGOs interviewed, the Digital Services Act is a very good opportunity to achieve a competitive Internal Market in the EU for digital consumer services.

Finally, reforms of **law enforcement action regarding illegal content online across borders** have to be launched as well according to another NGO. Most online platforms work in several Member States and have to face and comply with different jurisdictions. Some of the national laws they are asked to comply with may fail to respect human rights law, especially when talking about online content restrictions. In addition, problems can arise in circumstances where the substantive law on a free speech issue (e.g. holocaust denial) differs from country to country across the EU. The NGO added that online platforms often 'solve' these difficulties by applying their Terms of Service/Terms of Use rather than the national law, a practice that often leads to the blocking of content that is legal across the EU, or at least in some of its Member States. To avoid this, it was suggested that efficient cross-border mechanisms that respect the legal principles and safeguards of judicial cooperation should be established.

Regarding the existence of different content moderation practices in EU Member States, almost all NGOs, industry/trade associations and online platforms interviewed believed that **these differences hinder the fight against illegal content**. An NGO mentioned that both national and European regulatory proposals create a **complex patchwork of rules** as they usually involve different obligations in terms of content qualification, timeframes, proactive and filtering measures, sanctions, and reporting duties. On the European regulatory side, for example, initiatives to combat hate speech tend to mandate or encourage the removal of content within 24 hours, whilst the TERREG Proposal would require the removal of terrorist content within an hour. On the national side, several NGOs considered that French and German initiatives against online hate speech are examples of initiatives adding another layer of complexity for online platforms operating in the EU.

It was also mentioned by a few of the NGOs and online platforms interviewed that different and diverging legal regimes **increase compliance costs while also being a source of legal uncertainty**

[110] According to Article 3 of the ECD, online service providers in the EU are only subject to the rules of their country of origin or home country, i.e. the country where they are established.

regarding the qualification of content as illegal and the scope of responsibilities and obligations of online platforms. It was mentioned by a trade association that developing moderation tools and rules country-by-country "*will not help European platform operators to scale, but instead will result in gradual or fragmented investment in tools and processes, and ultimately less effective action*".

Several online platforms stressed that **harmonised and transparent 'notice-and-action' processes, which are implemented coherently throughout the EU Member States** are key for facilitating the fight against illegal content. A harmonised approach would enable online platforms to have more clarity on what they must do to fulfil their legal responsibilities, while upholding fundamental rights. EU-wide rules would also help prevent competition distortions and remove obstacles to the free movement of information society services.

Despite the clear wish of most stakeholders for an EU harmonised approach to content moderation of illegal content online, a few of the NGOs and industry/trade associations interviewed stressed the **difficulties linked to harmonisation of the rules at EU level**. Harmonisation is in practice difficult as there are different definitions of what is illegal depending on the Member State. National laws also vary considerably in how they restrict freedom of expression. For example, some Member States criminalise content such as blasphemy, while others have abrogated blasphemy laws. Another example is that many Member States prohibit hate speech but apply those prohibitions differently based on the cultural and historical context of their particular country. Additionally, given the cultural and language differences among the EU Member States, several NGOs acknowledged that it is necessary that automated content moderation tools respect these dissimilarities. It was suggested that algorithms of these tools should be trained with datasets relevant for certain regions and groups of society and should be applied locally.

3.4. Other issues

3.4.1. Liability under the e-Commerce Directive

The question on the current EU liability regime of Internet intermediaries enshrined in the e-Commerce Directive (ECD) has resulted in a high rate of stakeholders' responses. Most of the online platforms (i.e. seven out of the nine platforms interviewed) considered the terms of the **existing liability principles of intermediary service providers as fit-for-purpose**. This is because it allows for protection of fundamental rights, the rule of law and the open Internet. Furthermore, most of the platforms stated that the 'notice-and-takedown' regime enshrined in Articles 12 to 14 and the no-general monitoring obligation principle under Article 15 of the ECD should remain the "*cornerstone of online enforcement*". Consequently, from the majority of online platforms' perspective, this regime is usually the fairest and most effective way to tackle illegal content online issues.

However, two online platforms indicated that the **concept of 'active' and 'passive' intermediary service providers should be reformed**. This is because this concept may no longer adequately reflect the economic, social, and technical reality of current services across their lifecycle. Consequently, many services find themselves out of the scope of the existing legal framework or are unsure of the applicable EU legal regime in question. In addition, one of the online platforms pointed out that although the current EU liability regime is "*still highly relevant, and future-proof*", it should be "*strengthened and reinforced*". This could especially be achieved by clarifying the applicability of intermediary liability exemptions, since there is "*a significant fragmentation and diverging national interpretations of Articles 12-15, as well as ambiguities in the e-Commerce Directive*".

In comparison to the online platforms, half of the interviewed NGOs stated that the limited liability provisions have proven to encourage innovation and entrepreneurship in the Digital Single Market.

Consequently, according to half of the NGOs at stake, the liability exemption should be maintained, while the other half of the NGOs pointed out that the current EU liability principles of Internet intermediaries have become insufficient (both due to the development of the online platforms and to technological advancements). Thus, they would need to be adjusted carefully, in order to take into account the diversity of online platforms, and to distinguish between purely content hosting platforms and those that actively curate content via algorithmic ordering or other significant mediation.

Regarding the concept of increased liability of good faith actors, two stakeholders, namely one platform and one NGO, referred in their answers to a **'Good Samaritan' clause** that is enshrined in US law[111]. These two stakeholders disagreed between themselves upon the reasonableness of potential introduction of such a clause into EU law. The online platform at stake stated that the main area of reform should be to address the increased liability of good faith actors and that the 'Good Samaritan' clause could encourage more online platforms to strengthen their moderation capabilities. On the other hand, the NGO emphasised that the clause in question is a US legal concept that is not fit-for-purpose in EU law. This is because it would not help to address the challenges for platforms' responsibility for cyber safety and consumer protection, while encouraging platforms to play a role of deciders on what content should be allowed.

3.4.2. Freedom of speech issues

In the context of an online content moderation practice which should be implemented to respect the freedom of expression and information, five of the online platforms that we interviewed acknowledged the **need for moderation measures that are tailored to the different online services**. It has been indicated that such an approach (focusing on adjusting a content moderation mechanism to various online services) could ensure the most effective response while safeguarding freedom of expression and information. Considering the fundamental freedoms, the interviewed online platforms expressed different views on the content moderation mechanism in question. One of the platforms suggested to implement certain moderation standards. These standards should require platforms to deploy "*clear and accessible*":

- policies adequate for the specific service;
- reporting procedures concerning service misuse; and
- effective proceedings with such reports.

This platform has also stated that the system of 'notice-and-takedown' should be maintained, while online platforms should refrain from the use of mandating filters or automated technologies. Additionally, another platform pointed out the importance of users' access to appeal requests and platforms' transparency reports.

According to another two online platforms, the future EU legislative instruments should **avoid establishing incentives for undertakings to remove more content online than necessary**, for instance by enforcing too short timeframes for illegal content online removal and high fines for non-removal of user-generated items. Consequently, these two platforms acknowledged that platforms' incentive to over-remove legal content constitutes the most considerable threat of unjustified interference with fundamental rights.

[111] Section 230(c) of the US Communication Decency Act states that: "*(…) No provider or user of an interactive computer service shall be held liable on account of any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable (…)*".

Regarding the mechanism related to the respect of the freedom of expression and information, five out of eight NGOs agreed that **online platforms should establish relevant internal mechanisms**. These procedures will ensure that platforms reach moderation decisions in a non-arbitrary and transparent manner. The building blocks of these platforms' internal mechanisms should, according to the NGOs in question, comprise of:

- transparency reports published on a regular basis and including data on posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines;
- 'notice-and-takedown' processes including a human review that are fair and transparent; and
- appeal procedures providing a meaningful option for timely appeal regarding any illegal content online removal or account suspension.

Furthermore, one of the NGOs stated that the recommended scheme should be the **removal of illegal content at source following a Court order**. This would prevent law enforcement authorities (i.e. administrative authorities), companies, trusted reporters[112] (e.g. NGOs) or users to play roles of arbitrators of illegality. Another NGO suggested to establish at EU level an independent consultative 'Observatory on Freedom of Speech and Information'. This independent body could serve to defend the principles regarding human rights enshrined in the *Convention for the Protection of Human Rights and Fundamental Freedoms* and in the *Charter of Fundamental Rights of the European Union*.

3.4.3. Online discrimination issues

Five of out nine online platforms interviewed shared their views on safeguarding the fundamental right not to be discriminated against. Two of the platforms and two NGOs emphasised the role of **education** of Internet users. These interviewees indicated that in order to reduce new forms of online discrimination, online users and consumers must learn how to tackle prejudices. One of the NGOs highlighted that it is crucial for Internet users to have "*a strong critical sense*" that would allow them to better understand the source of information. In the context of hampering the development of new forms of online discrimination, one of the online platforms has launched a programme aiming to better understand the issues concerning hate speech and to prepare an effective and proportionate reaction to it. This programme's objective is to provide online users with "*a positive alternative to violence and extremism*". The programme encompasses campaigns responding to far-right extremism, Islamist terrorism, and disinformation and conspiracy theories posted by extremist/hate groups.

In order to address the issue of online discrimination, another online platform has updated their existing Community Standards/Guidelines on hate speech. These Guidelines clearly indicate that videos alleging a superiority of a given group or justifying religion, age, race, gender, caste, or sexual discrimination, segregation or exclusion, are prohibited.

One online platform and one NGO pointed out that there are **certain user groups, which may experience disproportionate impact of incorrect content moderation decisions**. According to the aforementioned interviewees, such a problem is **exacerbated** by platforms' intensifying reliance on flawed **automated content filtering technologies** to accomplish content moderation at scale. Importantly, the aforementioned NGO stated that machine-learning algorithms "*learn about the world from their training data, they copy and can further amplify social bias already reflected in the society*". Consequently, algorithms can become biased based on elements such as ethnicity, political affiliation, gender, language dialects and other cultural differences. Therefore, as it has been indicated by one of the NGOs, online platforms should not only cooperate more with expert flaggers (such as anti-

[112] Note that trusted flaggers are private individuals, while NGOs and other civil society organisations are treated as trusted reporters.

discrimination groups, anti-racism groups, etc.), but also ensure that content moderation staff "*receive sufficient training to be able to identify and appropriately deal with discriminatory content*".

3.5. Specific private initiatives

3.5.1. Facebook: Oversight Board for content moderation decisions

In September 2019, Facebook proposed to establish by the summer 2020 an **independent Oversight Board to ensure a fair and independent decision-making on Facebook's content moderation practice**[113]. This Oversight Board should be funded by a USD 130 million (estimated EUR 118 million) trust fund that is completely independent of Facebook and cannot be revoked for a period of at least 6 years[114]. It will be composed of up to 40 members whose identity is made public with varied and diversified skills, knowledge and expertise while ensuring geographical representativeness and appointed for a remunerated term of three years renewable twice[115]. The four co-chairs of the Oversight Board have been selected directly by Facebook while the other members will be designated on the basis of their qualifications by a committee of the Oversight Board and appointed by the trustees, after selection by Facebook and the co-chairs[116]. A dedicated staff will support the members of the Oversight Board[117].

The Oversight Board will provide policy guidance/advisory opinion on Facebook's content policies. More importantly, the Oversight Board will also review specific content moderation cases that could be submitted by Facebook or its users once the Facebook internal recourses have been exhausted[118]. The Oversight Board is free to decide which cases it reviews but must refrain from reviewing a case if its decision is likely to result in criminal liability or regulatory sanctions. A panel of five members takes the decision where at least one should come from the region concerned by the case[119]. The panel should obtain from Facebook the information necessary to decide the case and may receive written statements from the content author or complainant. It may also gather information (from experts or otherwise) necessary to provide context[120]. The panel should review the cases on the basis of Facebook's content policies and values while taking into account the human rights standards that protect freedom of expression[121]. Decisions taken by the Oversight Board are binding on Facebook[122], made public and clearly justified[123].

While this is a step in the right direction, **some commentators have proposed improvements to the Facebook proposal**. *Article 19* (see Section 3.5.2. below) calls for better transparency and improvements in internal removal procedures before the establishment of the Oversight Board. It also fears that the global level at which the Board operates will make it difficult to understand local contexts

---

113 The website of the Oversight Board is available at: https://www.oversightboard.com/.
114 Oversight Board Bylaws, Article 2, Section 1.3.1., available at: https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf.
For more information about the Trust, see Oversight Board Bylaws, Article 4.
115 Oversight Board Charter, Article 1, available at: https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf.
116 Oversight Board Charter, Article 1. The four co-chairs are Catalina Botero-Marino (former special rapporteur on freedom of expression of the Organization of the American States), Jamal Greene (law professor at Columbia), Michael W. McConnell (law professor at Stanford) and Helle Thorning-Schmidt (former prime minister of Denmark).
117 Thomas Hugues (former Executive Director for *Article 19*) has been appointed in January 2020 by Facebook as the first Director of the administration staff, see https://about.fb.com/news/2020/01/facebooks-oversight-board/.
118 Oversight Board Charter, Article 2. See also bylaws, Article 1, Section 3.
119 Oversight Board Charter, Article 3. See also bylaws, Article 1, Section 3.1.3.
120 Oversight Board Charter, Article 3.
121 Oversight Board Charter, Article 2.
122 Oversight Board Charter, Article 4.
123 Oversight Board Charter, Article 3.

(social, political, cultural, historical, linguistic, etc.) and their complexity[124]. Latonero also notes that the selection of the Oversight Board members by Facebook may undermine their independence and that the decisions will be made based on Facebook's values and content policies and not merely based on the international human rights standards[125]. More radically, Gosh claims that existing problems with online content are not related to poor moderation but to the very business model of several online platforms in terms of consumer engagement, constant online advertising offers, massive collection of personal data and the use of various sophisticated algorithms[126].

3.5.2. *Article 19*: Social Media Council initiative

Article 19 (2018), an advocacy charity defending freedom of expression, observes that the initiatives proposed by the online platforms to moderate online content in general lack transparency, do not offer satisfactory remedies or procedural safeguards to users, and do not sufficiently protect freedom of expression and other fundamental rights. *Article 19* also notes that the Terms of Service/Terms of Use and Community Standards/Guidelines of the online platforms generally restrict more the freedom of expression than the international fundamental rights standards, in particular because they are based on the lowest common denominator between the different national legislations applicable to content.

To solve some of these issues, *Article 19* proposes the establishment of a **Social Media Council (SMC) to provide an open, transparent, participatory, independent and accountable forum to review moderation practices**[127]. The SMC will gather social media platforms, medias, journalists, bloggers, academics, civil society organisations and any other stakeholder. This mechanism would be based on international standards of human rights but without creating legal obligations. The SMC will have a case review role, which will possibly be complemented with more general advisory roles[128].

Regarding geographic scope, *Article 19* envisages a network of councils set up at national level that would be governed by a global code of principles based on international fundamental rights standards, while applying them according to the local context. Other options were also discussed such as a regional or global SMC or a hybrid model combining a global Council with a network of national Councils[129]. In any case, *Article 19* underlines the importance of the national/local level to ensure a good understanding of the context (cultural, social, political, historical, linguistic, religious, etc.) of the content moderation dispute to be decided. Thus, the initiative proposed by *Article 19* could

---

[124] *Article 19*, "Facebook: New oversight board is not sufficient to safeguard freedom of expression online", 18 September 2019, available at: https://www.article19.org/resources/facebook-new-oversight-board-is-not-sufficient-to-safeguard-freedom-of-expression-online/.

[125] Latonero, Can Facebook's Oversight Board Win People's Trust?, 2020, available at: https://hbr.org/2020/01/can-facebooks-oversight-board-win-peoples-trust. Currently, in the bylaws, a somewhat flawed wording refers to fundamental rights, stating that the Oversight Board "*will be guided by relevant human rights principles and [...] [will provide] analysis of how the board's decisions have considered or tracked the international human rights implicated by a case*".

[126] D. Ghosh, Facebook's Oversight Board Is Not Enough, 2019, available at: https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough.

[127] See Docquir, The Social Media Council: Bringing Human Rights Standards to Content Moderation on Social Media, 2019, available at: https://www.cigionline.org/articles/social-media-council-bringing-human-rights-standards-content-moderation-social-media. *Article 19* is currently working on the preparation of a document bringing together the relevant recommendations for the establishment of the SMC and has obtained funding from the Open Society Foundation and is considering the possibility of launching a pilot experiment of SMC in a European country.

[128] One of the main criticisms of a pure case review function was the potentially very high number of requests for review, which seems very delicate for a structure such as SMC. It was then proposed that only the most important cases or by group of content types should be submitted to SMC for review: Conference Report of February 2019, "Social Media Councils: From Concepts to Reality", organised by Stanford University's Global Digital Policy Incubator, *Article 19* and the UN Special Rapporteur on the Right to Freedom of Opinion and Expression, available at: https://cyber.fsi.stanford.edu/gdpi/content/social-media-councils-concept-reality-conference-report, pp. 12-13.

[129] Each model has its advantages and disadvantages. The national SMC takes maximum account of the realities and needs of local contexts, there are risks of interference or even appropriation by States as well as difficulties in determining the jurisdiction of a given country. The global SMC would in principle make decisions more uniform but would not take into account local contexts and would have difficulties in ensuring diverse and pluralistic representation.

complement Facebook's Oversight Board, in particular to provide the necessary insights and understanding of local contexts.

3.5.3. Twitter: BlueSky initiative to build decentralised standards for social networks

In December 2019, Twitter announced the funding of an independent research group to develop **decentralised standards for social networks that can be used by different content moderation providers**[130]. Compared to the initiatives of Facebook or *Article 19* (which consist in reviewing/evaluating online platforms' content moderation practices), the Twitter initiative offers a different solution. Such standards could both reduce criticism of the platforms' content moderation practices and provide opportunities for new competitors, as control of content would no longer be concentrated in the hands of a few dominant companies. It also has the advantage of trying to offset the dominance and influence of the major platforms on online expression[131].

3.6. Specific practices during the COVID-19 pandemic

3.6.1. Specific measures to tackle illegal content online

The COVID-19 outbreak led to an increase of illegal content online such as scams, misleading advertising, and unfair, misleading and deceptive business practices by rogue traders. A **Common Position of the Consumer Protection Cooperation Network (CPC) and the European Commission requires the online platforms to better identify, remove and prevent the reappearance of such types of illegal content**[132]. In addition, the European Commission and the CPC notify the online platforms of breaches of EU consumer law most often committed in the context of COVID-19, such as unsupported claims that products prevent or cure the coronavirus (without strong scientific evidence or which are inconsistent with official experts' opinion), pressure selling techniques and excessive pricing.

On that basis, **many online platforms have set up specific and privileged communication channels to strengthen cooperation with national consumer protection authorities** to enable them to report illegal practices[133]. They have also implemented **a series of measures** such as automated and human monitoring of content based on keywords and product categories at risk of scam, introduction of algorithms and proactive measures to combat price fraud, prohibition on the sale or advertising of certain products (e.g. masks and disinfectant gel), information campaigns for both consumers and sellers. The first results of these measures are positive with a high number of content removal (up to 1

---

130 K. Paul and M. Vengattil, "Twitter plans to build 'decentralized standard' for social networks", 11 December 2019, available at: https://www.reuters.com/article/us-twitter-content/twitter-plans-to-build-decentralized-standard-for-social-networks-idUSKBN1YF2EN.

131 *Article 19*, "Why decentralisation of content moderation might be the best way to protect freedom of expression online", 30 March 2020, available at: https://www.article19.org/resources/why-decentralisation-of-content-moderation-might-be-the-best-way-to-protect-freedom-of-expression-online/.
See also R. Price, "Twitter is trying to build a new decentralized social media service that could transform its business – or present new kinds of headaches", 11 December 2019, available at: https://www.businessinsider.fr/us/bluesky-twitter-team-decentralised-social-media-2019-12.

132 Common position of European Commission and Consumer Protection Cooperation Network 20 March 2020 on stopping scams and tackling unfair business practices on online platforms in the context of the Coronavirus outbreak in the EU, available at: https://ec.europa.eu/info/sites/info/files/live_work_travel_in_the_eu/consumers/documents/cpc_common_position_covid19.pdf.
On the basis of this common position, the Commissioner for Justice and Consumers has written to various industry players (online platforms, social media, search engines and online marketplaces) to ask them to cooperate in removing these coronavirus-related scams and unfair practices. See European Commission, "Scams related to COVID-19", available at: https://ec.europa.eu/info/live-work-travel-eu/consumers/enforcement-consumer-protection/scams-related-covid-19_en.

133 Such as Alibaba, Allegro, Amazon, Cdiscount, eBay, Facebook, Google, Microsoft (Bing), Rakuten, Verizon Media (Yahoo) and Wish. Their answers are available at: https://ec.europa.eu/info/publications/detailed-replies-provided-platforms_en.

million per week for the largest online platforms) and with hundreds of thousands of price frauds detected[134].

3.6.2. Specific measures to tackle online disinformation

During the COVID-19 crisis, **several online platforms increased their effort to fight disinformation and implemented measures related to content removal, promotion of credible, authoritative and relevant information**, reducing the dissemination of some content, and reducing the possibilities of online advertising. Table 4 below contains an exemplary and non-exhaustive list of measures adopted by online platforms during the pandemic[135].

Table 4: Online content moderation practices in times of COVID-19

Online platforms Measures	Facebook apps[136]	YouTube - Google	Twitter[137]
Content removal	Removal of erroneous information that could lead to imminent physical damage and of false claims regarding treatment, availability of essential services, location or severity of the pandemic. Removal of content flagged as erroneous by fact-checkers and limits placed on forwarding possibilities on WhatsApp.	Removal of content discouraging people from treatment or claiming that dangerous substances are healthy.	Removal of content containing a clear call to action likely to pose a risk for health or well-being, denial of established scientific facts, spamming behaviour.
Promotion of credible, authoritative and reliable information	Coronavirus Information Centre at the newsfeed's top with updates from WHO and national health authorities.	Priority given to authoritative information (e.g. WHO, Centres for Disease Control and Prevention) from the homepages and in searches.	Prioritisation of credible, authoritative content at the top of search and verification of accounts providing credible updates about COVID-19.
Advertising prohibitions	Advertising relating to necessary health care equipment linked to COVID-19 (masks, hands sanitiser, COVID-19 tests) or to products guaranteeing cure or protection against COVID-19.	On Google Ads, blocking of advertising capitalising on COVID-19.	List of restrictions established (e.g. promotion of some products related to COVID-19 as necessary, health care equipment linked to COVID-19, distasteful references to COVID-19, sensational content or likely to incite panic, price inflation, etc.).

[134] European Commission, Summary of platforms' measures, available at: https://ec.europa.eu/info/sites/info/files/live_work_travel_in_the_eu/consumers/documents/summaryofresponses_update_08042020.pdf.

[135] See also the Joint industry statement of 17 March 2020 of Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter and Youtube on working together to combat misinformation.

[136] In an update, Facebook gives an overview of the measures put in place in Europe and around the world to counter misinformation about the pandemic and harmful content on its apps (Facebook, Messenger, Instagram and WhatsApp). See Facebook, "Keeping People Safe and Informed About the Coronavirus", available at: https://about.fb.com/news/2020/04/coronavirus/. The social network also indicates its willingness to work on the issue with the European institutions, Member States, WHO and other industry players.

[137] See Twitter, "Coronavirus: Staying safe and informed on Twitter", available at: https://blog.twitter.com/en_us/topics/company/2020/covid-19.html.

Advertising demonetisation	Covered indirectly by advertising prohibitions.	Reduced opportunities to monetise videos that make more than one reference to COVID-19, except for certain actors (e.g. news media).	Reduced opportunities to monetise content that make direct or indirect reference to COVID-19.
Free advertising or ad credits for WHO, CDC, national health authorities, etc.	Yes.	Yes.	Information not available.
Support (incl. financial) for news industry and fact-checkers	Yes.	Yes.	Yes.

Source: Authors' own elaboration

3.6.3. Results from the platforms interviews

Five out of nine interviewed online platforms responded to the question regarding any specific content moderation practices implemented during the COVID-19 pandemic to deal with fake news and unsafe products.

According to the first and second of these online platforms, as a consequence of the measures in question aimed at responding to COVID-19 and at prioritising the wellbeing of platforms' employees and extended workforce, the platforms started **relying more on technology to assist with some of the work normally done by human reviewers**. This means that automated systems will start removing some content without human review, allowing the platforms to continue to act quickly to remove violative online content and protect their ecosystem. These actions may lead to users and creators experiencing increased video removal, including some videos that may not violate the platforms' policies. The platforms will not issue strikes on this content except in cases where they have high confidence that the content is violative. Creators who think that their content was removed in error are allowed to appeal the decision and a dedicated platforms' teams will review the decision. However, due to workforce precautions implemented by the platforms because of COVID-19, the process will result in delayed appeal reviews. The measures also help people stay informed and connected, and support small businesses and other organisations. Moreover, the second of the two online platforms at stake is providing easy access to authoritative information from health authorities alongside new data and visualisations. This new format allows for user-friendly navigation through information and resources, and it will make it possible to add more information over time as it becomes available. The second of these two online platforms also committed to be more cautious about what content is promoted, including livestreams.

The third online platform that has provided answers related to measures implemented in relation to COVID-19 stated that they have taken **precautionary steps, under their policy on disallowed content in advertising**. Their aim is to block ads that are directly related to COVID-19. This policy bans advertising on 'sensitive' issues, and the platform deploys this policy provision to prohibit all advertising exploiting COVID-19 for commercial gain, spreading misinformation, or that may pose a danger to users' health or safety. To this end, the platform implemented a complex ranking process across all search results, which puts emphasis both on data relevance and on ensuring that high-authority sources of information rank higher in search results than low-authority sites.

In this context, the third online platform deployed **additional measures aimed at promoting access to trusted information and to combat fraud and misinformation**. As also explained in Table 4, these measures comprise of ensuring that answers and helpful public authorities announcements concerning COVID-19 will be listed at the top of search results for a number of COVID-19-related search queries. Measures also include displaying (for many searches regarding COVID-19) task panes with credible information in main places on the first page of search results, such as the top right-hand side of the page. Moreover, the platform cooperates with a highly professional news rating service. This service operates a coronavirus misinformation tracker listing all of the news and information sites in France, Italy, Germany, the UK, and the US that it has identified as publishing materially false information about COVID-19. When users of the aforementioned online platform have installed the plug-in and navigate to these sites, a warning label will appear notifying the user that the information on the site is not reliable.

The remaining two online platforms, i.e. the fourth and the fifth, have indicated that in the light of the COVID-19 crisis, they have issued a global policy on COVID-19 content to provide guidance to their local moderation teams on how to treat COVID-19 related content on their platforms. In addition, the fourth platform stated that their Terms of Service/Terms of Use were amended to temporarily prohibit posting ads concerning particular goods as a response to local government policies. The fifth platform noted that their moderators remain highly vigilant for false information regarding COVID-19.

4. INTERNATIONAL BENCHMARKING

KEY FINDINGS

The analysis conducted in six countries/regions of the world shows that **all of these countries have a regulatory and policy framework related to illegal content online**. The policies may relate either to online hate speech, defamation, child sexual abuse material, copyright infringements or online disinformation. The large majority of the countries investigated have regulatory measures in place related to at least one or several of these areas. Some **policy guidelines have also been established** by some countries to guide online platforms when moderating illegal content online. In almost all countries covered, NGOs and academics have recommended ways to improve the online content moderation practices.

In addition to the regulatory and policy measures put in place to frame illegal content online, **most of the online platforms worldwide apply their own Terms of Service/Terms of Use to moderate online content**, which often stem from the large US online platforms.

In comparison to the set of measures identified in the various countries, the **EU seems to have one of the most developed regulatory frameworks related to illegal content online and its moderation by online platforms**.

On the international scale, the EU appears to have one of the most comprehensive regulatory frameworks for tackling illegal content online[138]. In other regions of the world, few regulations apply and the online content moderation practices are mainly based on the Terms of Service/Terms of Use of the online platforms themselves. The international online platforms often follow the American content moderation practices given that the largest online platforms originate from the US. In parallel with platforms' standardised practices, NGOs and content moderation platforms around the world suggest best practices and recommendations to better moderate illegal content online.

In this section, the regulatory framework and the best practices related to online content moderation in six countries/regions of the world, namely Australia, Latin American countries, Canada, China, Japan and the US, are investigated. These countries represent a sample of different jurisdictions from different continents with an important population, which imply a large number of Internet users and the need to regulate the content they post.

The objective of this section is to present best practices, which could serve as models to follow at EU level. For the purpose of this study, desk research was conducted and, wherever possible, complemented with information from the stakeholders interviewed. Three specific criteria were applied for the research (regulatory framework, policy framework and recommendations on best practices) and the analysis for each country is presented according to these criteria:

- **regulatory and policy framework:** in this first section, the legal and policy background of the country regarding moderation of online content is presented. The aim is to establish whether there are any relevant laws or policy measures regulating illegal content online or online content moderation. In the majority of countries where the research has been conducted, a regulatory framework related to illegal content online and online content moderation was

[138] This is one of the results of our analysis and it has also been mentioned by several stakeholders interviewed while conducting the study, notably by a few of the online platforms and an industry association.

found. In all countries investigated, it was also found that the online platforms mainly apply their own Terms of Service/Terms of Use to regulate online content. However, in some countries (such as Australia, Latin American countries, Canada, China, and Japan), specific guidelines or reports from national authorities or relevant stakeholders (content moderation services companies, copyright organisations) explain how online content moderation works; and

- **recommendations on best practices**: in this second section, some of the best practices recommended by academia and NGOs from the countries concerned are presented, on how to improve the content moderation mechanisms. In some countries, no specific best practices of these types were found (such as China and Japan), while in others, NGOs' website or academic papers indicated suggestions on how to better moderate online content (such as in Latin American countries, Australia, Canada, and the US).

4.1. United States

4.1.1. Regulatory and policy framework

In the US, the First Amendment of the American Constitution establishes the right to free speech for individuals and prevents the government from infringing on this right. The First Amendment, however, does not similarly bind online platforms. As a result, they are able to establish their own content policies and Codes of Conduct that often restrict speech that could not be prohibited by the government under the First Amendment[139]. For example, Facebook, and most recently Tumblr, prohibit the dissemination of adult content and graphic nudity on their platforms. Under the First Amendment, however, such speech prohibitions by the government would be unconstitutional.

The ability of online platforms to moderate content in the US comes from Section 230 of the **Communications Decency Act (CDA), which gives online platforms broad immunity from liability for user-generated content** posted on their websites[140]. The purpose of this grant of immunity was both to encourage platforms to be 'Good Samaritans' and take an active role in removing offensive content, and also to avoid free speech problems of collateral censorship[141].

One of the online platforms interviewed while conducting this research suggested that the current EU regime applicable to online platforms should get inspired by the immunity stemming from Section 230 of the CDA. According to the interviewee, the EU regime does not include a clear provision which would protect online platforms from liability should their proactive content monitoring prove imperfect. It was thus suggested that the EU regime should use an approach similar to the US one to minimise risks in this area and focus on 'Good Samaritan' principles – whereby platforms that take good-faith proactive action against illegal content are not deemed to have 'actual knowledge' of illegality as a result. The interviewee believed that this could remove negative incentives that prevent proactive measures today.

In addition, in August 2010 the US adopted the Speech Act, which shields US journalists, publishers (both print and online) and bloggers from foreign lawsuits[142].

---

139 Singh (2019).

140 Codified at Title 47, Section 230 of US Code, available at: https://www.law.cornell.edu/uscode/text/47/230.

141 See *Zeran v. America Online*, Incorporated, 129 F.3d 327, 330 (4th Circuit 1997), available at: https://www.eff.org/files/zeran-v-aol.pdf.

142 An act to amend title 28, US Code, to prohibit recognition and enforcement of foreign defamation judgments and certain foreign judgments against the providers of interactive computer services, August 2010, available at: https://www.gpo.gov/fdsys/pkg/PLAW-111publ223/html/PLAW-111publ223.htm.

In the field of **illegal content online of sexual character**, the mandate given by the US Congress to the **National Centre for Missing & Exploited Children (NCMEC)** can be a best practice. The US dropped from hosting 43% of such content in 2017 to 25% in 2018 and 2019[143]. Given this success, the European legislators could consider transposing the reporting obligation that applies to online platforms in the US[144]. The Children´s Internet Protection Act[145] requires that US schools and libraries adopt measures to protect children from harmful content (pornography and child pornography). The Law does not block content per se, but it requires these organisations to block the content as a condition for federal funding. The US considers the production, distribution and use of child pornography as a criminal offence[146].

The legal framework related to freedom of speech in the US and the protection offered to online platforms enable the platforms to **regulate illegal content online on their own, through the use of their Terms of Service/Terms of Use**[147].

4.1.2. Recommendations on best practices

Several American academic papers explained how the largest online platforms regulate online content in the context of the American legal system. Some researchers argue that "*to best understand online speech, we must abandon traditional doctrinal and regulatory analogies and understand these private content platforms as systems of governance*"[148]. This is interesting in order to understand how to best address the issue of illegal content online, and especially how regulation can interact with these online platforms to combat illegal content online.

The online platforms are seen by some American researchers as "*self-regulating*[149] *private entities, governing speech within the coverage of the First Amendment*[150] *by reflecting the democratic culture and norms of their users*"[151]. Researchers argue that the biggest threat this private system of governance poses to democratic culture is the "*loss of a fair opportunity to participate, which is compounded by the system's lack of direct accountability to its users*". Solutions to this issue could be simple changes to the architecture and governance systems put in place by the platforms. If this fails and that regulation is needed, some argue that it should be designed to strike a balance between preserving the democratising forces of the Internet and protecting the generative power of the online platforms, with a full and accurate understanding of how and why these platforms operate[152].

Several US scholars show that it is important to keep in mind how the online platforms are operating and to understand the specificities of these entities before implementing regulatory changes in the area of online content moderation.

In parallel with the standardised online content moderation practices put in place by US-based platforms, **some best practices on how to better moderate illegal content online** have been proposed by American NGOs and content moderation platforms. For example, the NGO *Anti-Defamation League* established best practices for responding to cyberhate, from the side of online

---

143 See page 26 of the INHOPE annual report published in 2019: https://bit.ly/2VN0dti.
144 Title 18 US Code, Section 2258A, Reporting requirements of providers, available at: https://www.law.cornell.edu/uscode/text/18/2258A.
145 Children's Internet Protection Act, available at: http://ifea.net/cipa.pdf.
146 Citizen's guide to US Federal law on child pornography, available at: https://www.justice.gov/criminal-ceos/citizens-guide-us-federal-law-child-pornography.
147 Samples (2019).
148 Klonick (2017).
149 See generally Freeman (2000) and Michael (1995).
150 See generally Schauer (2004).
151 See generally Balkin (1995).
152 Klonick (2017).

platforms and the Internet community[153]. On the side of online platforms, it recommends for example that platforms should take reports about cyberhate "*seriously, mindful of the fundamental principles of freedom of expression, human dignity, personal safety and respect for the rule of law*". The platforms should offer user-friendly mechanisms and procedures for reporting hateful content. They should also respond to user reports in a timely manner. Finally, they should enforce whatever sanctions their Terms of Service/Terms of Use contemplate "*in a consistent and fair manner*". On the side of the Internet community, the NGO advises that the Internet community "*should identify, implement and/or encourage effective strategies of counter-speech*", including direct response, comedy and satire when appropriate, or simply setting the record straight. The Internet community should share knowledge and help developing educational materials and programmes that encourage critical thinking in both proactive and reactive online activity.

Best practices on how to moderate and manage content on news and political websites were also proposed by a leading industry publication based in the US[154]. For example, they suggested that there is a need for "*human eyes*" reviewing the content, that the online platforms should coordinate with law enforcement officials and support social bureaus (such as suicide lines), and should build "*a cultural model that supports the brand values*".

4.2. Canada

4.2.1. Regulatory and policy framework

The Canadian government is "*actively considering*" regulating online platforms since 2019 as it believes that "*self-regulation of the platforms has failed*"[155]. Notably, on 30 January 2019 the government of **Canada announced a series of actions to strengthen** the Canadian electoral system facing October 2019 elections, including a call for action from online platforms to increase transparency and authenticity in their systems[156]. Hate speech in Canada is addressed in the Criminal Code (Revised Statutes of Canada, 1985, c. C-46) and is not specific to digital communications[157].

Regarding **online child sexual abuse material**, a specific law (Act respecting the mandatory reporting of Internet child pornography by persons who provide an Internet service, 2011) applies to online platforms[158]. It requires them to report to the *Canadian Centre for Child Protection* tip-offs they receive regarding websites where child pornography may be publicly available, to notify police and to safeguard evidence if they believe that a child pornography offence has been committed using an Internet service that they provide. The production, distribution or use of child pornography are criminal offences in the Criminal Code[159]. Regarding the limitation of liability of online platforms, it applies only in case of 'innocent dissemination'. This occurs if there is no actual (or supposed) knowledge of the defamation contained in the material being disseminated and no negligence in failing to know that

[153] Anti-Defamation League, *Best Practices for Responding to Cyberhate*. Available at: https://www.adl.org/best-practices-for-responding-to-cyberhate.

[154] See Woodul, *7 best practices for managing and moderating user content on news and political sites*, Social media today, 2011, available at: https://www.socialmediatoday.com/content/7-best-practices-managing-and-moderating-user-content-news-and-political-sites.

[155] See Boutilier, Oved and Silverman, Canadian government says it's considering regulating Facebook and other social media giants, 2019, available at: https://www.thespec.com/news/canada/2019/04/09/canadian-government-says-it-s-considering-regulating-facebook-and-other-social-media-giants.html.

[156] Initiatives of the Government of Canada related to democratic institutions, Government of Canada's website, available at: https://www.canada.ca/en/democratic-institutions.html.

[157] Canadian Criminal Code, Revised Statutes of Canada, 1985, c. C-46, available at: https://laws-lois.justice.gc.ca/PDF/C-46.pdf.

[158] An Act respecting the mandatory reporting of Internet child pornography by persons who provide an Internet service, Statutes of Canada, 2011, c. 4, Justice Law Canadian website, available at: https://laws-lois.justice.gc.ca/eng/annualstatutes/2011_4/FullText.html.

[159] Section 163.1., Canadian Criminal Code, available at: https://laws-lois.justice.gc.ca/eng/acts/C-46/section-163.1.html.

the material contained the defamation at the time of its dissemination. This position was confirmed in the Internet context in several Supreme Court decisions (example: Crookes vs. Newton, 2011)[160].

Concerning the Canadian policy framework related to online content moderation, in December 2018 the Standing Committee on Access to Information, Privacy and Ethics published a report, in which it acknowledged the nature and gravity of **online content moderation and spread of disinformation on online platforms**[161]. It also proposed several recommendations to improve Canada's response to the challenges. One of the recommendations was for example that the Government of Canada enact legislation to further regulate online platforms. This would impose certain obligations on online platforms regarding:

- the labelling of content produced algorithmically;
- the labelling of paid advertisement online;
- the removal of inauthentic and fraudulent accounts; and
- the removal of manifestly illegal content, such as hate speech.

The idea to provide the mandate (to an existing or a new regulatory body) to proactively audit algorithms was also put forward.

Despite several debates on the topic in the country, at the date of the present study no laws were passed in Canada to regulate online content moderation or online platforms. Hence, in Canada the online platforms are applying their own Terms of Service/Terms of Use to moderate content.

4.2.2. Recommendations on best practices

In Canada, academics from the University of British Columbia and Concordia University have prepared a report which outlines how governments and online platforms can better address hate and harassment. This is the initiative to improve the current regulatory approaches in Canada that cannot address the speed, scale and global reach of harmful speech on online platforms.

The list of recommended actions[162] that shall be taken by the government to improve the regulatory approach towards online content moderation include the creation of a task force to improve government enforcement of existing policies. It also includes ensuring platform transparency and launching a public process to develop responses to issues of harmful speech and online content moderation more broadly.

It also includes the development of a **multi-stakeholder Moderation Standards Council** to strengthen and coordinate action by online platforms and stakeholders. This Council would enable online platforms, civil society and stakeholders to meet public expectations and government requirements on content moderation. It would improve transparency and help online platforms develop and implement Codes of Conduct on addressing harmful speech. It would create an appeal process to address complaints about content moderation policies and decisions.

Finally, the report recommends the establishment of a **civil society capacity to address harmful speech online**. In the context of civil society, governments could assist in various ways such as, direct funding, indirect funding via research institutes and academic granting agencies, pressuring online

---

[160] *Crookes v. Newton*, Supreme Court of Canada, 2011, available at: https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/7963/index.do.

[161] Report of the Canadian Standing Committee on Access to Information, Privacy and Ethics, *Democracy under threat: risks and solutions in the era of disinformation and data monopoly*, December 2018, 42nd Parliament, 1st Session, available at: https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/report-17.

[162] Tenove, Tworek, McKelvey (2018).

platforms to share data or provide reports to civil society, and addressing liability concerns of online platforms. The recommended practices also include building capacity for research and monitoring campaigns of disinformation, computational propaganda and harmful speech or 'Election Contact Group' in Canada to improve communication between civil society and online platforms before and during elections.

4.3. Australia

4.3.1. Regulatory and policy framework

Regarding the regulatory framework, **Australia passed a law in 2019**[163] that requires online platforms to **"*expeditiously*" remove violent content** from their platforms in light of the Christchurch massacre in New Zealand. According to that law, social media executives are subject to jail sentence or their companies face fines if the content is not taken down "expeditiously".

As for the policy framework, Australia put in place tailor-made strategies and guidelines for moderating content submitted by a business' end-users. At present, online platforms such as Facebook, Twitter and Instagram are consistently finding ways on how to improve their existing moderation practices. Maintaining a solid plan and a flexible approach to checking user-generated content is important to prevent majorly alarming and damaging content from surfacing online[164].

Among the most common strategies mentioned by the Australian moderation services company *New Media Services Pty Ltd.*, are[165]:

- assigning moderators for different social media pages, accounts and community threads;
- adding content filters and keywords to automate which terms and phrases to ban;
- enabling member participation in moderating content, either through flagging or reporting spam and inappropriate comments;
- expanding the scope of moderation by allowing moderators to check direct messages, comments, reviews and reports sent directly by community members; and
- muting notorious spammers or disabling their account for days.

4.3.2. Recommendations on best practices

As discussed in the previous section, the **Australian company New Media Services has created a list of DO's and DON'Ts regarding online content moderation**[166]. On the one hand, the list presents the DO's of online content moderation stating, for example, that the nature of the business needs to be considered, as it will dictate what content should be allowed or denied. The company suggests also that the target audience needs to be known, as it helps create boundaries on what can be considered appropriate and inappropriate. It was also recommended that a platform's users should always be informed of any issues with moderated content to ensure they understand why one image gets approved while another does not.

---

[163] Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019, A Bill for an Act to amend the Criminal Code Act 1995, and for related purposes, available at: https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/s1201_first-senate/toc_pdf/1908121.pdf;fileType=application%2Fpdf.

[164] *The Fundamental Basics of Content Moderation,* by New Media Services Pty, 2019, available at: https://newmediaservices.com.au/2019/01/22/fundamental-basics-of-content-moderation/.

[165] *Ibidem*.

[166] *Content moderation DO's and DON'Ts,* by New Media Services Pty, 2017, available at: https://newmediaservices.com.au/content-moderation-dos-and-donts/.

On the other hand, the company proposes DON'Ts regarding online content moderation[167], for example, automated filters should not be too much relied on, given that "*programmes are only as smart as we make them*", and a person will always find a way to bypass auto filters (like profanity filters). The company also advised against being overly general in moderation practices and that not making moderation practices fit the audience affects business negatively. It also mentions the dangers of over-moderation and of under-moderation.

4.4. Latin American countries

4.4.1. Regulatory and policy framework

In Brazil, there are no specific laws regulating fake news or hate speech online. Former President Tamer vetoed the provisions of Law 13488, of 6 October 2017 (on electoral rules) which aimed to impose obligations on platforms to take down hate speech and fake news content against political parties and politicians during electoral campaigns[168]. Regarding online child sexual abuse material, the Federal Law 11,829/2008 amended the *Child and Adolescent Statute* to criminalise the production, reproduction, fixation (by any means), sale and distribution of abusive content (video or photo containing sex/nudity scenes/images of children/minors) on the Internet[169]. Internet law (12,965/14) determines that users have the right to exercise parental control over online content. Online platforms and public authorities must jointly promote the use of parental control tools and digital inclusion of children and teenagers (Article 29)[170]. In Brazil, there is no liability of Internet Service Providers (ISPs) for illegal content online but there is a liability for online service providers. Indeed, Internet Law[171] establishes subsidiary liability for hosting third party-generated content if a violation of intimacy rights (nudity/sex content, for example) is found (Article 21); the Civil Code establishes a general liability rule (Article 927) and indemnities also applied for defamation (Article 953)[172]; and the Criminal Code (Article 139) establishes imprisonment (3 months to 1 year) and a fine as penalties.

In Argentina, the National Institute Against Discrimination, Xenophobia and Racism has an observatory on discrimination on the Internet[173]. They have a mediation role between industry and citizens. The Institute gathers the queries and complaints made by citizens and contacts the social networks or other platforms to reach agreements that satisfy all the parties (for discrimination of any type, including cyberbullying). In Argentina, illegal content online can be taken down, if ordered by a Court. Enacom, an autonomous and decentralised entity that operates within the scope of the Head of the Cabinet of Ministers of the Nation, publishes Court decisions to block illegal websites[174]. According to a 2014 Supreme Court ruling, extrajudicial requests can take down content that is obviously illegal, while the remaining cases require an order of a competent (judicial or administrative) authority[175].

---

167 *Ibidem*.
168 Law 13.488 of 06 October 2017, available at: http://legis.senado.leg.br/norma/26248253.
169 Law 11.829 of 25 November 2008, available at: https://www2.camara.leg.br/legin/fed/lei/2008/lei-11829-25-novembro-2008-584363-norma-pl.html.
170 Law 12.965 of 23 April 2014, available at: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm.
171 *Ibidem*.
172 Law 10.406 of 10 January 2002, available at: http://www.planalto.gov.br/ccivil_03/leis/2002/L10406.htm.
173 Website of National Institute against Discrimination, Xenophobia and Racism, available at: https://www.argentina.gob.ar/inadi.
174 List of court orders for website blocking, on National Communication Entity's website, available at: https://www.enacom.gob.ar/bloqueo-de-sitios-web_p3286.
175 Rodríguez, María Belén c/Google Inc. and other s/damages, 2014, available at: http://www.saij.gob.ar/corte-suprema-justicia-nacion-federal-ciudad-autonoma-buenos-aires-rodriguez-maria-belen-google-inc-otro-danos-perjuicios-fa14000161-2014-10-28/123456789-161-0004-1ots-eupmocsollaf.

In Colombia, the anti-discrimination law 1482/2011[176], amended by law 1752/2015[177], punishes acts of discrimination based on race, ethnicity, religion, nationality, political or philosophical ideology, sex or sexual orientation, disability or other forms of discrimination. Penalties are increased if a conduct is carried out through the use of mass media (there is no definition of mass media but it could apply to widely used platforms). Regarding fake news shared online, the Colombian National Police implemented a strategy entitled 'TRUE' to prevent fake news attacks[178]. Concerning the child sexual abuse online material, the Law 679/2002 (fight against child abuse) establishes that all persons must prevent, block, combat and denounce exploitation, accommodation, use, publication, distribution of images, texts, documents, audio files, improper use of global information networks, or establishment of telematics links of any kind related to pornographic or alluding to sexual activities of minors material.

4.4.2. Recommendations on best practices

The Latin American Observatory for Regulation, Media and Convergence has proposed the **Latin American perspective for content moderation processes that are compatible with international fundamental rights standards**[179]. The document entitled "Contributions for the democratic regulation of big platforms to ensure freedom of expression online" contains a number of recommendations on specific principles, standards and measures designed to protect users' freedom of expression and guarantee a free and open Internet[180].

The recommendations concern a wide range of aspects, such as:

- the scope of content moderation;
- online platforms' Terms of Service/Terms of Use and conditions;
- transparency of the actions;
- application of online platforms' policies;
- the right to defence and repair; and
- the accountability of online platforms for content moderation.

Some of the recommendations include for example that platforms should directly incorporate into their Terms of Service/Terms of Use or Community Standards/Guidelines the relevant fundamental rights principles that ensure the measures related to the content will be guided by the same criteria governing the protection of freedom of expression by any means. These principles include transparency, accountability, due process, necessity, proportionality, non-discrimination and the right to defence and repair.

According to these recommendations, platforms should also ensure full respect for consumer rights and they should issue periodic transparency reports on the application of their Community Standards/Guidelines that include at least full data describing the categories of user content that are restricted, data on how many content moderation actions were initiated by a user's report (flag), a trusted flagger programme or by the proactive application of Community Standards/Guidelines (for

[176] Law 1482 of 2011 National Level, available at: http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=44932#1482.
[177] Law 1752 of 2015 National Level, available at: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=61858#1.
[178] National Police of the Republic of Colombia's website, Article *Policías vs* 'fake news', 2018, available at: https://www.policia.gov.co/noticia/policias-vs-fake-news.
[179] Latin American Observatory for Regulation, Media and Convergence's website, available at: https://www.observacom.org.
[180] *Contributions for the democratic regulation of big platforms to ensure freedom of expression online, A Latin American perspective for content moderation processes that are compatible with international human rights standards*, Latin American Observatory for Regulation, Media and Convergence. Available at: https://www.observacom.org/wp-content/uploads/2019/08/Contributions-for-the-democratic-regulation-of-big-platforms-to-ensure-freedom-of-expression-online.pdf.

example, through the use of a machine learning algorithm). It should also include the data on the number of decisions that were effectively appealed or determined to have been made in error and the data reflecting whether the company performs a proactive audit of its non-appealed moderation decisions, as well as the error rates that the company found.

4.5. China

4.5.1. Regulatory and policy framework

Regarding the Chinese regulatory system on online content moderation, according to a China-based company TechNode dealing with Chinese technology and start-up ecosystems, new online content regulations, namely 'regulations on ecological governance of online content' have been passed in December 2019[181]. As TechNode has indicated, the Chinese authorities (i.e. the Office of the Central Cyberspace Affairs Commission)[182] are likely to come down heavily on rule-breaking content after the March 2019 deadline and may suspend or shut down offending online platforms. The rules linked to online content moderation practices stemming from the analysed Chinese regulatory system encompass:

- the ban of exaggerated, rumour-laden, sexually provocative, and dangerous content which may incite copycats;
- the ban of acts which infringe on personal privacy, use of new technology to engage in illegal acts such as artificial intelligence-powered face swapping, buying traffic, and use of the Communist Party or state symbols in marketing campaigns;
- platforms using personalised recommendation algorithms must include controls for manual intervention and user choice;
- advertisements are considered online content; and
- platforms are encouraged to create content versions suitable for minors.

Regarding the policy framework, according to the Electronic Frontier Foundation, the Copyright Society of China launched in April 2017 its 12426 Copyright Monitoring *Centre*, which is dedicated to scanning the Chinese Internet for evidence of copyright infringement[183]. This Centre is said to be able to monitor video, music and images found on "*mainstream audio and video sites and graphic portals, small and medium vertical websites, community platforms, cloud and Peer-to-Peer sites, Smart TV, external set-top boxes, aggregation apps, and so on*"[184].

More precisely, when the 12426 Copyright Monitoring Centre finds content that matches material submitted to it by a copyright holder, the Centre provides them with a streamlined notification and takedown machine, from the issuance of warning notices through to the provision of mediation services. The Centre's technology service provider also provides platforms with filtering technology that can allow infringing materials to be blocked from upload or download to begin with, obviating the need for a separate takedown procedure[185].

---

[181] See Au, Online content rules leave platforms holding the bag, 2019, TechnNode, available at: https://technode.com/2019/12/23/online-content-rules-leave-platforms-holding-the-bag/.

[182] The Regulations are available in Chinese on the website of the Office of the Central Cyberspace Affairs Commission at: http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm.

[183] Malcolm, Chinese Government and Hollywood Launch Snoop-and-Censor Copyright Filter, Electronic Frontier Foundation. 2017, available at: https://www.eff.org/deeplinks/2017/04/chinese-snooping-foreshadows-future-copyright-enforcement.

[184] *Ibidem*.

[185] *Ibidem*.

4.5.2. Recommendations on best practices

No specific best practices recommended by NGOs, academia or other relevant stakeholders could be found in China.

4.6. Japan

4.6.1. Regulatory and policy framework

No specific legal framework related to online content moderation could be found in Japan.

Regarding the policy framework, in Japan the rules governing the online platform users are enshrined in their extensive Terms of Service/Terms of Use and include the regulation of adult content and Intellectual Property. Generally, the rules are enforced by a user-based reporting system and volunteer moderators who investigate complaints and impose sanctions where necessary[186]. The identified online content moderation practices stemming from the analysed Japanese system are:

- a user-based reporting system and volunteer moderators who investigate complaints and impose sanctions;
- sanctions for users which include temporary suspension of the user's privileges and in the most extreme cases can extend to the deletion of the user's profile;
- the content labelling system which is fundamental to the rules and indeed a failure to adequately label content is a breach of the rules in itself. The voluntary labelling is an effective mode of regulation and constitutes a potential source of community cohesion through neighbourly practices;
- posting works created entirely by others is not allowed;
- works that are created using references (including photos) must provide a link or citation to the original reference material;
- use of official art, trademarks, or copyrighted materials, such as corporate logos or commercial music, is not allowed; and
- the moderators' attempts to ensure that online content is consistent with the United States' censorship practices and conventions.

4.6.2. Recommendations on best practices

No specific best practices recommended by NGOs, academia or other relevant stakeholders could be found in Japan.

[186] Pearson, Giddens and Tranter (2018).

5. POLICY RECOMMENDATIONS FOR THE DIGITAL SERVICES ACT

KEY FINDINGS

The revised EU regulatory framework for online content moderation, which will result from the **forthcoming Digital Services Act, could be based on the following objectives and principles**:

- sufficient and effective safeguards to protect fundamental rights;
- a strengthening of the Digital Single Market;
- a level playing field between offline and online activities;
- technological neutrality;
- incentives for all stakeholders to minimise the risk of errors of over and under removal of content;
- proportionality of the potential negative impact of the content and the size of the platforms; and
- coherence with existing content-specific EU legislation.

The **baseline regulatory regime** applicable to all types of content and all categories of platforms could strengthen in an appropriate and proportionate manner the responsibility of the online platforms to ensure a safer Internet. To do that, it could include a **set of fully harmonised rules on procedural accountability** to allow public oversight of the way in which platforms moderate content. Those rules could include: (i) common EU principles to improve and harmonise the **'notice-and-takedown' procedure** to facilitate reporting by users; (ii) the encouragement for the platforms to take, where appropriate, proportionate, specific **proactive measures** including with automated means; and (iii) the strengthening of the **cooperation with public enforcement authorities**. Those new rules could be based on the measures recommended by the European Commission in its 2018 Recommendation on measures to effectively tackle illegal online content as well as on the measures imposed on Video-Sharing Platforms by the revised 2018 Audio-Visual Media Services Directive.

This baseline regulatory regime could be complemented with **stricter rules imposing more obligations, when the risk of online harm is higher**. Stricter rules are already imposed **according to the type of content**: more obligations are imposed for the moderation of the online content with the highest potential negative impact on the society such as terrorist content, child sexual abuse material, racist and xenophobic hate speech and some copyright violations. Stricter rules could also be imposed **according to the size of the platform**: more obligations could be imposed on the platforms whose number of users is above a certain threshold, which could be designated as Public Space Content-Sharing Platforms (PSCSPs).

As often in EU law, enforcement is the weak spot and therefore, the forthcoming Digital Services Act should ensure that any online content moderation rule is **enforced effectively**. Such enforcement should be ensured **by public authorities**, in particular regulatory authorities and judicial courts. The 'country of origin' principle should be maintained, hence the **online platforms should in principle be supervised by the authorities of the country where they are established**. However, the authorities of the country of establishment may not have sufficient means and incentives to supervise the largest platforms; hence, an **EU authority could be set up to supervise the PSCSPs**. In addition, the enforcement could be improved with, on the one hand, a **better coordination between national authorities** by relying on the Consumer Protection Cooperation Network and, on the other hand, **better information disclosure** in the context of Court proceedings.

Given the massive explosion of online content, **public authorities may not be sufficiently well-geared to ensure the enforcement of content moderation rules and may need to be complemented with private bodies**. Those could be the platforms themselves, self-regulatory bodies or co-regulatory bodies. The involvement of private bodies seems inevitable, but should not lead to full delegation of State sovereign power to private firms or a privatisation of the public interest, hence co-regulation could be an effective tool, preferred to self-regulation.

Next to specific obligations regarding the moderation of illegal content online, **complementary broader measures are also necessary** such as more transparency on the way moderation is done and support to journalists, Civil Society Organisation or NGOs, which contribute to the fight against illegal content.

Based on the results of the previous sections, this section makes proposals to improve the EU regulatory framework for the moderation of content online within the context of the forthcoming Digital Services Act and the expected revision of the e-Commerce Directive.

5.1. Principles on which a reform should be based

The EU regulatory framework for the moderation of content online should **protect effectively the victims of illegal content while guaranteeing an appropriate balance among fundamental rights**. This could be achieved by **efficiently sharing the responsibility** for the detection and the removal of illegal online content among the many actors involved in the diffusion of such material and evolving towards a system of 'cooperative responsibility'[187]. Indeed, all stakeholders (such as platforms, users, competent authorities, experts, CSOs, NGOs, trusted flaggers, fact-checkers, news media and journalists, and researchers) should be involved for effective long-term solutions.

The regulatory framework could be based on the **following principles**:

- provide sufficient and effective safeguards for **EU standards relating to all fundamental rights**, in particular, freedom of expression, the right to privacy, the prohibition of discrimination and the right to a fair trial/effective remedy;
- strengthen the **Internal Market and alleviate national regulatory fragmentation**; this requires confidence of the Member States (and their citizens) that the regulation in the country of establishment is sufficiently protective and effectively enforced; in turn, this requires, on the one hand, a harmonisation of the main rules aimed to protect users and, on the other hand, cooperation and mutual assistance between the competent authorities of the Member States in charge of enforcing the rules;
- ensure **a level-playing field** between online and offline activities and ensure that what is illegal offline is also illegal online; the rules should also be **technologically and business neutral** and not favour one technology or business model over others;
- provide to all stakeholders involved in the removal of illegal online content the **right incentives to minimise the risk of errors**, of type I errors (over-removal) and of type II errors (under-removal);

[187] As suggested by Helberger, Pierson and Poell (2018). Also Buiten, de Streel and Peitz (2020).

- **be proportionate**, which could lead to a differentiation of rules according to the type of content (and its potential negative impact on the society) and according to the size of platforms (and their means and societal reach); at the same time, the multi-layered regulatory framework to which differentiation leads should remain **coherent**;
- be **sufficiently general** to be easily adaptable to technology and business models, which evolve quickly and often in unpredictable ways; to ensure legal certainty, these general rules could then be clarified by the European Commission in delegated or implementing acts or interpretative guidance; and
- be **enforced effectively**, on the basis of a smart combination **of traditional State enforcement mechanisms** with administrative and judicial authorities and **alternative private enforcement mechanisms** such as self- and co-regulation and out-of-courts dispute resolution tools.

5.2. The baseline regime: strengthening procedural accountability of online platforms

The baseline liability regime contained in the e-Commerce Directive could be amended or replaced by a Regulation in order to strengthen in an appropriate and proportionate manner the responsibility of the online platforms to ensure a safer Internet. The new rules could include a **set of fully harmonised rules on procedural accountability to allow public oversight of the way in which platforms moderate content**. These rules could make sure that platforms abide by good governance rules and practices which reflect EU democratic and fundamental right values. They could ensure oversight of the policies, processes and tools put in place by platforms to ensure that illegal content is taken down where needed. To remain proportionate, the smaller platforms could need to abide by the same set of procedural rules, but tailored according to their size, type and reach.

Those rules on procedural accountability could relate to the 'notice-and-takedown' procedure to facilitate reporting by users, the possibility and the need to take proactive measures to facilitate platforms' detection and the cooperation with public enforcement authorities (see Sections 5.2.1. and 5.2.2.). They could be based on the measures recommended by the European Commission in its Recommendation on measures to effectively tackle illegal online content as well as on the measures imposed on Video-Sharing Platforms by the revised Audio-Visual Media Services Directive[188]. In other words, it could be an integration into the hard-law of some soft-law recommendations and an extension of the rules currently applicable to VSPs to all online content platforms.

Importantly, **fewer obligations should be imposed for harmful content** than for illegal content as freedom of speech needs to be preserved. Relevant measures could include: closing false accounts and fighting bots; promoting independent counter-speech, relevant, authentic and trustworthy content (e.g. from experts); encouraging the finding of alternative content on general interest content; strengthening transparency measures, media literacy and democracy education; and making available parental control tools and rating systems.

[188] European Commission Recommendation 2018/334, Points 5-28; AVMSD, Article 28b. In addition, our proposed reforms would also meet the Santa Clara Principles on Transparency and Accountability in Content Moderation: https://www.santaclaraprinciples.org/.

5.2.1. Increased role for users and trusted flaggers

The forthcoming DSA with the expected revision of the ECD could introduce more expansive rules on **transparency concerning content removal**, their processing, mistakes, actors and notifications[189]. Such rules could also ensure personalised explanations for affected users and audits for authorities or researchers[190].

Providers of hosting services could set up mechanisms for notices that are easy to access, user-friendly and allow for automated submission. The **'notice-and-takedown' system could be facilitated and based on common principles defined at EU level**[191]. Husovec (2018) suggests to legislate only on the essential requirements of the process, and then leave the details to the standardisation process at the European Standards Organisations (CEN, CENELEC and ETSI), which can better reflect industry-wide best practices in different areas. Such technical standards could then serve as a proof of the provider's best efforts to comply with the 'notice-and-takedown' system as diligently as possible[192]. Technical standardisation could better foresee and keep up with automation, new techniques used and other market developments.

To reduce the risks of type I errors (over-removal) and ensure an appropriate balance among fundamental rights, the platform could[193]:

- encourage **notices** which are sufficiently precise and adequately substantiated;
- when practical and proportionate, first inform the content provider of the intention to suspend access to the supposedly illegal material and the reason of such suspension and give the provider the possibility to contest such suspension by submitting a **'counter-notice'**; and
- the platform could only remove the material from all platforms active in the EU after having assessed in a diligent manner, on the basis of the information given, the validity and the relevance of this 'counter-notice'.

However, in exceptional circumstances, when the illegality is manifest and relates to serious criminal offences involving a threat to the life or safety of persons (such as terrorist content), content may be removed immediately.

Online platforms could also cooperate more closely with hotlines and **trusted flaggers** that could be designated on the basis of clear and objective criteria based on expertise. Such cooperation may lead to fast-track procedures for notices submitted by trusted flaggers[194].

5.2.2. Preventive measures

Online platforms could be **encouraged to take, where appropriate, proportionate and specific proactive measures** in respect of illegal online content, including with automated means[195]. However, some safeguards could be in place and such proactive measure could not lead to a general monitoring that should continue to be prohibited.

---

189 Some of these recommendations are also mentioned in de Streel and Husovec (2020).

190 As recommended by the High-Level Expert Group on Artificial Intelligence (2019) and by the European Parliament Resolution of 12 February 2020 on Automated decision-making processes: Ensuring consumer protection, and free movement of goods and services.

191 Also Husovec (2017), Sartor (2017).

192 This is similar to the so-called "New Approach" used by the EU since the eighties in the field of technical standardisation and product safety and security.

193 European Commission, Recommendation 2018/334, Points 5-13; AVMSD, Article 28b(3) (d)-(e).

194 European Commission, Recommendation 2018/334, Points 25-27.

195 European Commission, Recommendation 2018/334, Points 18-20.

A **'Good Samaritan' clause** could be affirmed explicitly to ensure that the online platforms taking on proactive measures are not treated in a less favourable way than the ones not taking these measures[196]. Such a 'Good Samaritan' clause could aid platforms when taking voluntary measures, by removing the risk of being sanctioned for under-removal.

Reliance on **automated detecting tools** by intermediaries or users could be encouraged as an effective detection means, provided some safeguards be in place. This is part of the wider debate on the EU Regulation of Artificial Intelligence (AI), which should be based on the application of six key requirements:

- human agency and oversight;
- technical robustness and safety;
- privacy and data governance;
- transparency, diversity, non-discrimination and fairness;
- societal and environmental wellbeing; and
- accountability[197].

It is also key to note that the explainability obligations already imposed by the GDPR and other recent EU laws apply to automated content moderation practice[198]. Moreover, there may be a need for the large online platforms (which have the data, the expertise and the financial means to develop automated techniques) to **share these technologies** with the small and medium-sized or new platforms[199].

5.3. Aligning responsibility with risks

In addition to reforming the baseline regime applicable to all categories of platforms and all types of content, stricter rules increasing the responsibility of the platforms should be imposed when the risks of online harms also increase[200]. **To reflect such risk-based approach, differentiation could be made** according to:

- the **type of online content: more extensive obligations could be imposed regarding the moderation of the illegal content with the highest negative impact on the society**. This is already the case today as stricter rules are imposed against terrorist content, child sexual abuse material, racist and xenophobic hate speech and some copyright violations. All those rules should, on the one hand, be coherent with each other and with the baseline regime and, on the other hand, provide sufficient and effective safeguards to ensure the appropriate balance among fundamental rights set by the Court of Justice of the EU and the European Court of Human Rights (ECHR).
- the **size of the online platform: more extensive obligations could be imposed on the platforms with the largest size**. Thanks to their innovation, some content-sharing platforms have become so large and so important in the life of citizens that they are not merely running

[196] Also in this sense, Sartor (2017:29). As already explained, the European Commission considers that the 'Good Samaritan' clause is already compatible with the e-Commerce Directive: Communication on tackling illegal online content, COM(2017), p.13.
[197] European Communication White Paper of 19 February 2020 on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65; High-Level Expert Group on Artificial Intelligence, Ethics Guidelines of 8 April 2019 for Trustworthy AI.
[198] On these obligations and their technical implementation, see Bibal, Lognoul, de Streel and Frenay (2020).
[199] European Commission, Recommendation 2018/334, Point 28.
[200] For a law and economics approach of the liability rules of online platforms, see Buiten, de Streel and Peitz (2020).

a private space but now hosting part of the public space[201]. Such differentiation by platforms size is already emerging in EU law but should be affirmed more clearly in the forthcoming DSA. In practice, platforms with a number of users above a certain threshold, which could be designated as **Public Space Content-Sharing Platforms** (PSCSPs), could be subject to more extensive procedural accountability obligations. They could also be required to adopt regular transparency reports explaining how they moderate content with clear and comparable statistics. Finally, to increase the incentive to comply with those rules, the liability exemption of the ECD could be conditioned, for the PSCSPs, to the compliance with the stricter procedural accountability obligations. In other words, if a PSCSP does not set up an appropriate 'notice-and-takedown' mechanism or does not take appropriate proactive measures, the platform would not be able to rely on the liability exemption provided in the ECD. In addition, as explained below, those PSCSP could also be subject to a differentiated oversight and be supervised by an EU authority and not the authority of the Member State where the PSCSP is established.

5.4. Improving the effectiveness of the monitoring and enforcement

Effective enforcement is key and is often of the main weakness of EU law. To adapt the famous quote of Bill Clinton's strategist James Carville to EU law, "it's enforcement, stupid!". To ensure more effective enforcement mechanisms, the DSA could smartly combine traditional enforcement mechanisms with public authorities (along with a differentiation between small and big platforms) with alternative enforcements tools with private bodies.

5.4.1. Enforcement with public authorities

a. Public enforcement by independent authorities

The online platforms should be **supervised by the authorities of the country where they are established** according to the 'country of origin' principle. These authorities should be fully independent given the importance of their role in upholding freedom of expression, media plurality and press freedom. Moreover, the **cooperation and mutual assistance between Member States**, in particular between the country of origin where the online platform is established and the country of destination where the platform is offering its services, should be strengthened.

However, the authorities of the country of establishment may not have the ability nor the incentive to regulate the largest online content platforms, i.e. the PSCSPs subject to stricter moderation obligations (see above). For those platforms, EU rules **could be enforced by an independent EU regulator** - in close partnership with the national regulatory authorities - which would be sufficiently well funded to also conduct investigations into the operation of platforms[202]. Moreover, the EU independent authority could also maintain a database of which national authority is in charge of which platform.

b. Private enforcement

Where a moderation practice breaches the rights of users in at least two EU countries other than the EU country where the infringement originated or for widespread infringements, the mechanism set up under the EU Consumer Protection Cooperation Network Regulation could

[201] As suggested by Smith (2020), those public space platforms should now be regulated according to public law values and not any more according to private law values.

[202] In that regard, the enforcement of financial regulation on systemic banks by Single Supervisory Mechanism within the European Central Bank is an interesting starting point: Council Regulation 1024/2013 of 15 October 2013 conferring specific tasks on the European Central Bank concerning policies relating to the prudential supervision of credit institutions, O.J. [2013] L 287/63.

come into play[203]. According to this Regulation, national authorities should give a coordinated response to cross border infringements of EU consumer protection legislation, through a network that has been established between them. As explained above (Section 3.6.1.), the Consumer Protection Cooperation Network adopted a common position on stopping scams and unfair business practices on online platforms in the context of the COVID-19 outbreak[204]. Progressively, the mechanism has been broadened to cover breaches of a wide range of EU legislative instruments, which are no longer necessarily linked to consumers per se, such as breaches of the AVMSD. Alternatively, the EU independent authority could be called on to coordinate national responses and/or to adopt a decision.

Moreover, users should always have **access to a Court or another relevant judicial authority** to defend their rights. In the case of illegal activities carried out through the platforms, the victims should be able to initiate civil or criminal proceedings against the authors of such activities, in order to get compensation. To facilitate the identification of the authors, the victims should be able to receive the necessary information which is available to the platforms. Information requirements are already imposed by the ECD[205] but, in some Member States, the implementation of those obligations does not allow the victims to get such identification data. Thus, those information disclosure requirements could be strengthened (or, at least, clarified) in full compliance with the fundamental rights (in particular the right to privacy and the presumption of innocence).

5.4.2. Enforcement with private bodies

a. Codes of Conduct, self- and co-regulatory bodies

Codes of Conduct should continue to be encouraged as they can be very useful in fast moving industries where the best manners to achieve regulatory goals set in the law are not easy to determine. However, given their increasing importance, the DSA could impose **additional safeguards** on the manner such Codes are established and monitored under in order to increase their legitimacy, their effectiveness and compliance with fundamental rights, thus leading to a co-regulatory approach. In particular, the DSA could impose, on the one hand, that the Code of Conducts should be accepted by the main actors representing different interests at stake and, on the other hand, that their implementation should be regularly monitored independently with transparent and robust methodologies[206].

Moreover, as in the German NetzDG, the possibility of using a **self-regulatory body, recognised by the State to rule on the illegality of online content** (when it is not obviously illegal) could be explored in order to alleviate the risk of over-removal. This mechanism has only just been put in place in Germany so there are still lessons to be learned. However, the approach is attractive as it could discharge platforms from taking difficult decisions, while giving users certain safeguards and alleviating the possible incentives of platforms to over-removal out of fear of heavy fines and therefore prefer to remove content that is legal in case of doubt.

b. Out-of-Court dispute resolutions mechanisms

Dispute resolution is of fundamental importance as users need to be able to challenge decisions by platforms which may affect fundamental rights. **Access to dispute resolution should be made as**

---

[203] Regulation 2017/2394 of the European Parliament and of the Council of 12 December 2017 on cooperation between national authorities responsible for the enforcement of consumer protection laws and repealing Regulation 2006/2004, OJ [2017] L 345/1.

[204] The common position is available at: https://ec.europa.eu/info/sites/info/files/live_work_travel_in_the_eu/consumers/documents/cpc_common_position_covid19.pdf.

[205] ECD, Article 5. Also Directive 2004/48 of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights, O.J. [2004] L 195/16, Article 8.

[206] See AVMSD, new Article 4a introduced by Directive 2018/1808.

simple as possible, which is why Alternative Dispute Resolution (ADR) systems should be available in the country and language of where the alleged victim is located. These ADR systems should be **independent and well-funded** and provide for rapid, effective and impartial relief. In that regard, Fiala and Husovec (2018) propose to create an external ADR, which would be financed by higher fees paid by providers which erroneously take down the content and lower fees by users who complain without success. Such fees are meant to incentivise providers to improve their internal processes and provide a credible remedy to users to get their content reinstated and be heard by an impartial body.

5.5. Complementary measures

5.5.1. Transparency

Transparency obligations contribute to effective moderation of illegal and also harmful content **while safeguarding fundamental rights**. They inform users on key issues (such as origin of the content, identity of the author, possible sponsorship, amount paid to prioritise content, etc.) without affecting the content as such. They provide users with important contextual elements enabling them to assess the content they are confronted with. They help, for example, to detect online disinformation while respecting freedom of expression. Moreover, there is also a need to improve the **transparency of online advertising** (origin, identity of the sponsor, amounts received, etc.).

5.5.2. Supporting and empowering journalists and news media

Journalists and the news media also have a key role as they offer a wide range of information to develop critical thinking skills and equip citizens to detect manipulation and assess the illegality or harmful nature of online content. Subject to their own deontology rules, they provide independent, verified and objective information while cross-checking their sources. This allows, on the one hand, to raise public awareness of issues of general interest such as racism, xenophobia and terrorism and, on the other hand, to counter-balance and dilute online disinformation by providing deontological quality information as a counter-discourse. Thus, journalists and news media should be supported to enable them to propose a diversified and pluralistic offer of information. In addition, content moderation practices should not affect the journalistic content disseminated on online platforms.

5.5.3. Civil Society Organisations/NGOs and research/academic institutions

Civil Society Organisations and NGOs also contribute to the fight against illegal and harmful online content. They are developing media literacy, citizenship and democracy education actions as well as initiatives to develop critical thinking skills. Similarly, the **academic and research community** is essential to understand illegal and harmful online content, their origin, the identity of their authors, the reasons for their actions and to develop inclusive, innovative and effective solutions that respect fundamental rights.

REFERENCES

- Article 19 (2018), *Side-stepping rights: Regulating speech by contract*, Policy Brief, available at: https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf.
- Balkin, J. M. (1995), *Populism and Progressivism as Constitutional Categories*, 104 Yale Law Journal 1935.
- Bibal, A., Lognoul, M., De Streel, A. and Frenay, B. (2020), *Implementing Legal Requirements on Explainability in Machine Learning*, Artificial Intelligence and Law 28, forthcoming.
- Buiten, M., De Streel, A. and M. Peitz (2020), *Rethinking liability rules for online hosting platforms,* International Journal of Law and Information Technology 28, forthcoming.
- Coche E. (2018), *Privatised enforcement and the right to freedom of expression in a world confronted with terrorism propaganda online*, Internet Policy Review, 7(4).
- De Streel, A., Husovec, M., *The e-Commerce Directive as the cornerstone of the Internal Market: Assessment and Options for Reforms*, In-Depth Analysis the committee on the Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2020.
- ERGA (2020), *Report on disinformation: Assessment of the implementation of the Code of Practice*, available at: https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf.
- Fiala, L., Husovec, M. (2018), *Using Experimental Evidence to Design Optimal Notice and Takedown Process*, TILEC Discussion Paper 2018-028.
- Floridi L., Taddeo M. (2017), *The responsibility of Online Service Providers*, Springer.
- Freeman J. (2000), *The Private Role in Public Governance*, 75 New York University Law Review 543.
- Helberger, N., Pierson, J., Poell, T. (2018), *Governing online platforms: From contested to cooperative responsibility*, The Information Society 34(1), pp. 1-14.
- High-Level Group on Fake News and Online Disinformation (2018), *A multi-dimensional approach to disinformation*, available at: https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.
- Holznagel, B. (2018), *La loi d'application sur les réseaux – L'approche allemande pour lutter contre les "fausses nouvelles"*, in: Sauvageau F., Thibault S. & Trudel P. (dir.), *Les fausses nouvelles: nouveaux visages, nouveaux défis*, Presses de l'Université de Laval, pp. 197-214.
- Husovec, M. (2017), *Injunctions Against Intermediaries in the European Union: Accountable But Not Liable?*, Cambridge University Press.
- Husovec M. (2018), *The Promises of Algorithmic Copyright Enforcement: Takedown or Staydown? Which is Superior? And Why?*, 42(1) Columbia Journal of Law & the Arts, pp. 53-84.
- ICF, Grimaldi Studio Legale and 21c Consultancy (2018), *Overview of the legal framework of notice-and-action procedures in Member States*, Study for the European Commission.

- INHOPE, *Annual Report 2018*, available at: https://www.inhope.org/media/pages/the-facts/download-our-whitepapers/3976156299-1576235919/2019.12.13_ih_annual_report_digital.pdf.
- Jenay P. (2015), *Study on combating child sexual abuse online*, Study for the European Parliament.
- Klonick K. (2017), *The New Governors: The People, Rules, and Processes Governing Online Speech*", 131 Harvard Law Review 1598.
- Klonick K. (2019), *Does Facebook's Oversight Board Finally Solve the Problem of Online Speech?*, available at: https://www.cigionline.org/articles/does-facebooks-oversight-board-finally-solve-problem-online-speech?gclid=EAIaIQobChMIkKKvjLPT6AIVheh3Ch05JAGPEAAYASAAEgLVSfD_BwE.
- Kuczerawy A. (2018), *Intermediary Liability and Freedom of Expression in the EU: from Concepts to Safeguards*, Intersentia.
- Kukliš L. (2020), *Video-Sharing Platforms in AVMSD – A new kind of content regulation*, Research Handbook on EU Media Law and Policy, Elgar Publishing.
- McGonagle T. (2013), *The Council of Europe against online hate speech: Conundrums and challenges*, available at: https://rm.coe.int/16800c170f.
- Michael, D. C. (1995), *Federal Agency Use of Audited Self-Regulation as a Regulatory Technique*, 47 Administrative Law Review 171.
- Nash, V. (2019). *Revise and resubmit? Reviewing the 2019 Online Harms White Paper*, Journal of Media Law, 11:1, pp. 18-27.
- Nordermann, J. B., *Liability of Online Service Providers for Copyrighted Content – Regulatory Action Needed?*, In-Depth Analysis for the committee on the Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2018.
- Pearson, A., Giddens T. and Tranter K. (2018), *Law and Justice in Japanese Popular Culture: From Crime Fighting Robots to Duelling Pocket Monsters*, Routledge.
- Pierrat, E., Ullern, C. (2019), *Lutte contre la haine sur internet: quelle(s) décision(s) attendre du législateur?*, *Revue Lamy Droit de l'Immatériel* 160, pp. 50-52.
- Quintais, J. P. (2020), *The New Copyright in the Digital Single Market Directive: A Critical Look*, European Intellectual Property Review, 1.
- Quintel, T., Ullrich, C. (2019), *Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond*, in B. Petkova and T. Ojanen., *Fundamental Rights Protection Online: The Future Regulation of Intermediaries*, Edward Elgar.
- Ramboll (2018), *Evaluation of the implementation of the Alliance to Better Protect Minors Online*, Study for the European Commission.
- Rambaud, R. (2019), *Lutter contre la manipulation de l'information*, AJDA, 453.
- Samples, J. (2019), *Why the Government Should Not Regulate Content Moderation of Social Media*, available at: https://www.cato.org/publications/policy-analysis/why-government-should-not-regulate-content-moderation-social-media.

- Sartor, G., *Providers Liability: From the eCommerce Directive to the future*, In-Depth Analysis for the committee on the Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2017.
- Schauer, F. (2004), *The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience*, 117 Harvard Law Review 176.
- Seng, D. (2015), '*Who Watches the Watchmen?' An Empirical Analysis of Errors in DMCA Takedown Notices*', available at: SSRN.
- Singh, S. (2019), *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User Generated Content*.
- Smith, M., *Enforcement and cooperation between Member States: E-Commerce and the future Digital Services Act*, In-Depth Analysis Study for the committee on the Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2020.
- Tambini, D. (2019), *Reducing Online Harms through a Differentiated Duty of Care: A Response to the Online Harms White Paper*, available at: https://www.fljs.org/sites/www.fljs.org/files/publications/Reducing%20Online%20Harms%20through%20a%20Differentiated%20Duty%20of%20Care.pdf.
- Tenove C., Tworek H. J. S. and McKelvey F. (2018), *Poisoning Democracy: How Canada Can Address Harmful Speech Online,* The Public Policy Forum, available at: https://ppforum.ca/publications/poisoning-democracy-what-can-be-done-about-harmful-speech-online/.
- TILT (2016), *Role of online intermediaries: Summary of the public consultation*, Study for the European Commission.
- Tworek, H., Leerssen, P. (2019), *An analysis of Germany's NetzDG Law*, available at: https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.
- Urban, J. M., Karaganis, J., Schofield, B. L. (2017a), *Notice and Takedown: Online service provider and rightholder accounts of everyday practices*, Journal of Copyright Society 64, pp. 371-410.
- Urban, J. M., Schofield, B. L., Karaganis, J. (2017b), *Takedown in Two Worlds: An Empirical Analysis*, Journal of Copyright Society 64, pp. 483-520.
- Valcke, P. (2019), *The EU regulatory framework applicable to Audiovisual Media Services*. In Garzaniti, L., O'Regan, M., Valcke, P., De Streel, A. (eds.), *Telecommunications, Broadcasting and the Internet*, EU Competition Law & Regulation, 4th ed., Sweet & Maxwell.
- Van Eecke, P. (2011), *Online Service Providers and Liability: A Plea for a Balanced Approach*, Common Market Law Review 48, pp. 1455-1502.
- Van Hoboken, J. (2009), *Legal Space for Innovative Ordering: On the Need to Update Selection Intermediary Liability in the EU*, 13 International Journal of Communications Law & Policy.
- Van Hoboken, J., Pedro Quintais, J., Poort J. and Van Eijk N. (2018), *Hosting Intermediary Services and Illegal Content Online*, Study for the European Commission.
- Venturini and al. (2016), *Terms of Service and Human Rights: an Analysis of Online Platform Contracts*, Council of Europe, available at: https://bibliotecadigital.fgv.br/dspace/handle/10438/18231.

- VVA (2020), *Assessment of the implementation of the Code of Practice on Disinformation*, Study for the European Commission.
- Weinand, J. (2018), *Implementing the EU Audiovisual Media Services Directive*, Nomos.
- Wood, L., Perrin, W. (2019), *Online harm reduction – a statutory duty of care and a regulator*, Carnegie UK Trust, available at: https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf.

ANNEX I: ANALYSIS OF NATIONAL LAWS AND POLICIES ON ONLINE ILLEGAL AND HARMFUL CONTENT MODERATION

1. GERMANY: NETWORK ENFORCEMENT ACT (NETZDG)

The German *Network Enforcement Act* (NetzDG) was adopted in June 2017 to improve the enforcement of existing criminal provisions on the Internet and, more specifically, on social networks [207].

1.1. Scope of application

The NetzDG applies to social networks with more than 2 million registered users in Germany[208]. Two types of online platforms are excluded from the scope: the online platforms intended for the dissemination of specific contents (such as online gaming platforms, professional social networks, and online sales platforms) and the online platforms for journalistic and editorial contents[209].

In order to determine what is covered by illegal content, the NetzDG refers to 22 offences of the German Criminal Code[210] which includes child sexual abuse material, illegal hate speech (xenophobic and racist) and other types of hate speech, terrorist content, content infringing Intellectual Property Rights or online disinformation.

1.2. Obligations imposed on online platforms

With regard to the reporting of illegal content, online platforms must put in place an effective and transparent procedure that is easily recognisable, directly accessible and permanently available to users[211]. Such procedure must ensure that the online platform immediately becomes aware of the complaint submitted, that it analyses the legal or illegal nature of the litigious content and the possibility of its removal or blocking[212]. Moreover, in the event of the removal of illegal content, the online platform must ensure that it is secured and kept for evidence purposes for ten weeks[213].

After receiving a complaint about an alleged illegal content, the online platform has seven days to remove or block it. This time limit can be exceeded in two cases: (i) when the illegality of a content depends on the veracity of a factual allegation or other identifiable factual circumstances as the online platform may give the user the opportunity to comment on the complaint before taking a decision; (ii)

---

207 The NetzDG, is available at https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html. A new bill is currently proposed to amend the NetzDG. The Minister of Justice introduced a draft law in December 2019 to combat right-wing extremism and hate crime. It will necessarily have an impact on the NetzDG, particularly with a view to impose on social networks an obligation to transmit all content and usage data of users who are the subject of a complaint to the Federal Criminal Police Office with a possibility of fines up to EUR 50 million. See: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RefE_BekaempfungHatespeech.pdf;jsessionid=4E74DD0B0713364DD9F4C4DCDEB72109.2_cid289?__blob=publicationFile&v=1.

208 NetzDG, Article 1, Section 1 (2).

209 NetzDG, Article 1, Section 1 (1). It should be noted, however, that the situation of journalistic or editorial content that would be disseminated, not on journalistic and editorial content platforms, but on social networks remains uncertain. See Holznagel (2018).

210 NetzDG, Article 1, Section 1 (3). See also Tworek & Leerssen (2019). These illegal contents refers to: Dissemination of propaganda material of unconstitutional organisations and use of symbols of such organisations; Terrorist acts and formation of a terrorist organisation; Preparation or instructions for the commission of a serious act of violence that endangers the State; Counterfeiting (of objects as well as of information, news or factual assertions which, if believed, could have an influence on external security or on Germany's relations with other States); Public incitement to commit criminal offences, disturbance of the public peace by the threat of criminal offences or reward and approval of criminal offences; Incitement to hatred, violence or arbitrary measures according to the national, racial, religious or ethnic membership as well as outrages upon human dignity by insult, malicious denigration or defamation; Insulting denominations, religious communities and ideological associations; Representation/depiction of violence; Child pornography and making pornographic content accessible to minors; Defamation, insult and threat; Violation of intimate privacy through the taking of pictures; Falsification of evidentiary data.

211 NetzDG, Article 1, Section 3 (1).

212 NetzDG, Article 1, Section 3 (2).

213 NetzDG, Article 1, Section 3 (4).

when the online platform requires an *ad hoc* regulated self-regulatory body to rule on the illegality of a content[214]. The time limit is reduced to 24 hours in case of obviously illegal content (unless longer period is granted by the competent law enforcement authorities)[215].

Online platforms are also subject to transparency obligations and are obliged to inform and justify their decisions without delay to the complainant and the content's author[216]. Moreover, platforms that receive more than 100 complaints a year must publish a report detailing how they deal with them every six months[217]. Online platforms are also obliged to designate a representative in Germany to receive notifications or requests for information[218].

When an online platform fails to comply with those obligations, it faces an administrative fine, between EUR 500,000 to EUR 5 million depending on the type of infringement[219]. If the administrative authority justifies the infringement on the grounds that illegal content has not been blocked or removed, it must first obtain a judicial decision on the illegality[220].

1.3. Assessment

Holznagel (2018) and Tworek & Leerssen (2019) criticised the NetzDG for its impact on fundamental rights, in particular the freedom of expression. First, the scope of illegal content subject to the new obligations is too broad and could have been limited to content which denies the right of persons to exist or likely to lead to a disturbance of the public peace. Second, the law may lead to over-blocking as the sanctions are asymmetric, the online platforms being fined if they maintain illegal content but not when they remove legal content. The short timing under which online platforms should act exacerbates this perverse effect. Third, the NetzDG provides few redress mechanisms for the author of allegedly illegal content to complain. Although the author is informed about the online platform's decision regarding the content, the online platform has no obligation to allow them to express their point of view.

2. FRANCE: AVIA LAW ON ONLINE HATE SPEECH

The French Parliament has just adopted in May 2020 the law to fight hate speech over the Internet, the so-called Avia law[221].

2.1. Scope of application

The law applies to online platforms sharing public content and search engines whose activity in France exceeds a threshold that will be determined by decree, regardless of where they are established[222]. The

[214] NetzDG, Article 1, Section 3 (2) point 3.

[215] NetzDG, Article 1, Section 3 (2) point 2. In the sense of the NetzDG, the obvious illegality of a content is met when it can be identified as such within 24 hours by online platform employees qualified for this task.

[216] NetzDG, Article 1, Section 3 (5).

[217] NetzDG, Article 1, Section 2 (1). It should be noted that (2) of the same provision lists the various items of information that must be included in this report, including: the number of complaints received per year relating to illegal content; information relating in particular to the organisation and staff assigned to deal with complaints; a general description of the prevention efforts implemented by the platform; a description of the processes for transmitting complaints and the criteria for the removal or blocking of illegal content; figures reflecting the action taken on complaints (withdrawal, blocking, contestation, etc.); a report on the response times of the platforms to remove or block the reported illegal content; the measures put in place by the social network to inform both the complainant and the user affected by the removal or blockage of the illegal content; etc.

[218] NetzDG, Article 1, Section 5.

[219] NetzDG, Article 1, Section 4 (1) and (2).

[220] NetzDG, Article 1, Section 4 (5).

[221] French draft law to fight hate content on the Internet, as adopted in final reading by the National Assembly, adopted text n° 419, 13 May 2020. The law will enter into force on 1 July 2020. However, the text of the law has not yet been published in the Official Journal of the French Republic. Note that, on 18 May 2020, more than sixty senators appealed against the law to the French Constitutional Council.

[222] French draft law, Article 1.

law imposes a range of measures to combat the dissemination of illegal content online such as hate speech, child sexual abuse material and terrorist content[223].

2.2. Obligations

The French law imposes the following obligations:

- the strengthening of transparency at all levels, such as regarding processing of notifications, decisions taken and their justification, removal of content, sanctions and penalties, modalities applied to moderate illegal content online, technological and human means implemented by online platforms, internal appeal mechanisms, Community Standards/Guidelines, for advertisers which have commercial relationship with online platforms that have been subject of removal measures and also for minors[224];
- the implementation of a single notification system for illegal content on all online platforms that is directly accessible and easy-to-use[225]; on the basis of a notification system, online platforms and search engines will be obliged to remove any "manifestly illegal content" within of 24 hours; for terrorist content and child sexual abuse material, the law is stricter by requiring the removal within one hour[226]; and
- a series on enforcement mechanisms: the establishment of internal appeal mechanisms[227]; the strengthening of cooperation with judicial authorities[228], the imposition of heavy sanctions[229], the reinforcement of the powers of the High Audio-Visual Council[230], and the creation of a Hate Observatory[231].

2.3. Evaluation by the Commission

Contrary to the NetzDG, the European Commission issued critical observations against the French bill, pointing at risks of incompatibility with the ECD[232]. The Commission noted a risk of infringement of the 'country of origin' principle and a restriction of the freedom to provide information society services as a result of the obligations imposed to online platforms established outside France.

The Commission also mentioned a risk of incompatibility with the liability regime for intermediaries (in particular Articles 14 and 15 of the ECD). In its view, the drastic reduction of the information to be provided when sending a notification of illegal content to online platforms would not allow the

---

223 The offences concerned are those listed in Law 2004-575 on confidence in the digital economy: glorification of war crimes, of crimes against humanity, of certain other crimes and of terrorism, incitement to commit acts of terrorism, child pornography and offering pornographic content to minors, insults or incitement to hate based on a ground relating to race, religion, ethnicity, sex, sexual orientation or identity, gender identity, disability, negationism, sexual harassment, incitement to violence (including sexual and gender-based violence) and attacks on human dignity.

224 French draft law, Articles 1, 4, 5 and 9.

225 French draft law, Article 4. Article 2 of the French law also specifies the content and mandatory information to be included in notifications, depending on the status of the notifier.

226 This measure has been hotly debated, including by senators, who believe that it is inconsistent with the TERREG Proposal: Public Sénat, "Loi Avia: le Sénat rétablit sa version du texte et supprime le délit de "non-retrait" pour les plateformes Internet", 26 February 2020, available at https://www.publicsenat.fr/article/parlementaire/loi-avia-le-senat-retablit-sa-version-du-texte-et-supprime-le-delit-de-non.

227 French draft law, Article 4.

228 French draft law, Articles 1, 5 and 8. A whole series of measures are imposed by French law to strengthen cooperation with judicial authorities. For example, online platforms and search engines must temporarily keep removed content for judicial authorities and they must designate a representative on French territory. The judicial authorities that have imposed the removal of a content may also authorise an administrative authority to impose measures to prevent access to websites that re-broadcast such content.

229 French draft law, Articles 1, 6 and 7.

230 French draft law, Article 7. The High Audio-Visual Council is responsible for ensuring that online platforms and search engines respect their obligations. It may impose specific measures or penalties on them. The High Audio-Visual Council must also publish an annual report on the measures implemented and their effectiveness.

231 French draft law, Article 16.

232 Commission Decision of 22 November 2019, Notification 2019/412/F, Loi visant à lutter contre les contenus haineux sur internet: Emission d'observations prévues à l'article 5, paragraphe 2, de la directive 2015/1535, C(2019) 8585.

presumption of knowledge of the presence of illegal content and would constitute an "insufficiently precise and substantiated notification". The Commission also stated that the ECD does not preclude national legislators from requiring online platforms established on their national territory to act within a specified period. However, such a time limit must be proportionate and reasonable and must allow for more flexibility in justified situations (in particular, where the illegal nature of a content would require a more consequent assessment). Thus, the Commission considers that the imposition of the 24-hour time limit combined with heavy sanctions in case of non-compliance could create a disproportionate burden on online platforms and, in certain circumstances, a risk of excessive removal of content, thus infringing freedom of expression. Furthermore, the obligation to implement appropriate means to prevent the re-dissemination of deleted content is intended to impose a general obligation of monitoring on platforms, which is prohibited by Article 15 of the ECD. Indeed, online platforms may have to put in place automatic and generalised filtering of all their contents (including those that would require an in-depth assessment of their contexts to discover the illegality) in order to comply with this obligation.

3. FRANCE: LAWS ON INFORMATION MANIPULATION

In December 2018, two (one ordinary and one organic) laws relating to information manipulation were adopted to fight against the spread of false information during election periods[233].

3.1. Scope of application

The French laws refer to "false information" without giving a clear definition, although in the provision establishing the possibility of making an application to the judge hearing the summary proceedings ("*juge des référés*"), false information is defined as "inaccurate or misleading allegations or imputations of a fact likely to alter the truthfulness of the forthcoming election [...] disseminated in a deliberate, artificial or automated and massive manner through an online public communication service"[234]. The Constitutional Council decided that, to comply with respect the freedoms of expression and communication, the inaccurate or misleading nature of such allegations must be manifest and could not refer to opinions, parodies, partial inaccuracies or mere exaggerations and that the risk of altering the truthfulness of the election must also be manifest[235].

3.2. Obligations

The French laws introduce a new provision in the French Electoral Code imposing the following information requirements on online platforms whose activity exceeds 5 million unique visitors per month in France[236]: clear, fair and transparent information on the identity of the persons who pay them in order to promote certain information contents participating in the general interest debate and on the use of personal data in the promotion of such content; above a certain threshold, these online platforms must also make public the amounts of remuneration received for the promotion of content. Those obligations apply for the three months preceding the elections[237]. The platforms fulfil their

[233] Loi organique 2018-1201 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information, *J.O.R.F.*, 23 décembre 2018 and Loi 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information, *J.O.R.F.*, 23 décembre 2018. Both laws have been validated by the Constitutional Council with a narrow interpretation of the laws.

[234] Code électoral français, Article L163-2. Authors' own personal translation.

[235] Décision 2018-773 DC du Conseil constitutionnel du 20 décembre 2018, points 22-23.

[236] See Code électoral français, Article L163-1. See also Décret n° 2019-297 du 10 avril 2019 relatif aux obligations d'informations des opérateurs de plateforme en ligne assurant la promotion de contenus d'information se rattachant à un débat d'intérêt général, *J.O.R.F.*, 11 avril 2019.

[237] Code électoral français, Article L163-1.

transparency obligations through a register that is kept up to date and made publicly and electronically available[238]. If they fail to do so, they may be fined up to EUR 75,000[239].

The new provisions of the French Electoral Code allow the Public Prosecutor's Office, election candidates, various political parties, and any person with an interest, to submit an application to the judge hearing the summary proceedings. The aim of such action is that hosting platforms or, failing that, Internet service providers take proportionate and necessary measures to stop the dissemination of false information[240].

Online platforms whose activity exceeds a threshold of number of visits in France are also required to take several measures to fight the dissemination of false information likely to disturb public order or to distort the truthfulness of an election[241]. On the one hand, the French legislator imposes the introduction of a visible and easily accessible mechanism enabling Internet users to report false information[242]. On the other hand, the legislator foresees a non-exhaustive list of additional measures relating to the nature, origin and methods of dissemination of content, the fight against accounts disseminating massively false information, transparency of the algorithms used, the provision of information on the identity of persons whose contents participating in the general interest debate is promoted in return for payment and promotion of contents from media companies and news agencies and media literacy[243].

Online platforms must report annually to the French media regulator (the High Audiovisual Council) on the measures implemented[244] and must also appoint a legal representative to act as a point of contact on French territory[245].

In addition, if online platforms use algorithms to reference, classify or recommend information content participating in the general debate, they are subject to an obligation to publish aggregated statistics on the operation of the algorithms[246]. These statistics should indicate, for each content, the percentage of accesses not influenced by the recommendation, ranking or referencing algorithms and the percentage of accesses influenced by such algorithms.

The powers of the media regulator to contribute to the fight against the dissemination of false information that could disturb public order or undermine the elections have also been strengthened. The High Audiovisual Council must ensure compliance with the obligations imposed on online platforms and may make recommendations[247]. It is thus a co-regulatory mechanism that leaves the online platforms free to choose how to implement the measures to be taken, while forcing them to report to the media regulator.

4. UNITED KINGDOM: ONLINE HARMS WHITE PAPER

In April 2019 and on the basis of the work of Wood and Perrin (2019), the *Online Harms White Paper* was adopted with proposals to reduce online harms[248].

---

238 Code électoral français, Article L163-1.
239 Loi 2018-1202, Article 1.
240 Code électoral français, Article L163-2.
241 Loi 2018-1202, Article 11.
242 Loi 2018-1202, Article 11, al. 2.
243 Loi 2018-1202, Article 11, al. 3.
244 Loi 2018-1202, Article 11, al. 4.
245 Loi 2018-1202, Article 13.
246 Loi 2018-1202, Article 14.
247 Loi 2018-1202, Article 12.
248 Online Harms White Paper, available at https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper. The White Paper was opened for public consultation from April 2019 to July 2019. After receiving responses from a wide range of

4.1. Scope of application

The proposals apply to online platforms that allow users to share or discover user-generated content or interact with each other online[249]. The White Paper covers a wide range of online harms which can be sorted in three categories:

- harms with a clear definition (such as child sexual abuse material, terrorist content, organised immigration crime, modern slavery, extreme pornography, revenge porn, hate crim);
- harms with a less clear definition (such as cyberbullying, extremist content and activity, disinformation); and
- underage exposure to legal content (such as children accessing pornography).

However, the White Paper does not contain a clear definition of illegal content nor harmful content[250].

4.2. Obligations: A statutory duty of care

The White Paper recommends to impose a new statutory duty of care, under which online platforms would need to show how they take care of their users. Online platforms would be required to explicitly state what content and behaviour are acceptable on their sites and enforce this consistently and transparently. In practice, online platforms would have to determine and establish appropriate systems and processes to react to concerns over harmful and illegal content such as effective internal complaint mechanisms, transparent decision-making over actions taken in response to reports of harm; and relevant Terms of Service/Terms of Use[251].

The White Paper promotes a proportionate and risk-based approach[252]. This means that the regulator should focus on those online platforms that "pose the biggest and clearest risk of harm to users, either because of the scale of the platforms or because of known issues with serious harms"[253]. Moreover, differentiated expectations would be established on online platforms, depending on the legality of content and online platforms would not be forced to remove specific harmful content. Furthermore, specific rules would be applied to two types of illegal content (child sexual abuse material and terrorist content)[254]. Finally, the UK government also intends to introduce this legislation proportionately, minimising the regulatory burden on small online platforms[255].

The regulator would set out how to fulfil this duty of care in Codes of Practice. If online platforms want to fulfil this duty in a manner not set out in the Codes, they would have to explain and justify to the regulator how their alternative approach will effectively deliver the same or greater level of impact. An independent regulator would thus assess the compliance with this duty of care. The regulator is not defined yet, but most probably would be the OFCOM (the regulator and competition authority for the

---

respondents and having undertaken engagements with representatives from industry, civil society and others, the UK Government published its initial response in February 2020 and the full response is expected later this spring.

249 In practice, this definition captures social media platforms, cloud hosting providers, file hosting sites, public discussion forums, retailers who allow users to review products online, messaging services and search engines. In its initial response, the Government promises that the regulator will provide guidance to help companies understand whether or not they would fall into the scope of the Regulation. See the initial consultation response: https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response.

250 Online Harms White Paper, para. 2.2.

251 For a detailed list see Online Harms White Paper, para. 7.4.

252 Online Harms White Paper, para. 5.3.

253 Online Harms White Paper, para. 31 (Executive summary).

254 In this regard, the White Paper states that: "*Companies will be required to take particularly robust action to tackle terrorist use of the internet and online CSEA [Child Sexual Exploitation and Abuse]. The government will have the power to issue directions to the regulator regarding the content of the codes of practice for these harms, and will also approve the draft codes before they are brought into effect. Similarly, the regulator will not normally agree to companies adopting proposals which diverge from these two codes of practice, and will require a high burden of proof that alternative proposals will be effective*". Online Harms White Paper, para. 3.10.

255 Online Harms White Paper, para. 4.5.

UK communications and media industries). The regulator would therefore have the power to take action against companies that do not meet their duty of care (e.g. issuing substantial fines, imposing liability on individual members of senior management, etc.).

4.3. Assessment

Wood and Perrin (2019) note that the ECD permits duties of care introduced by Member States and that the AVMSD already requires Member States to take some form of regulatory action in relation to Video-Sharing Platforms. However, many commentators criticise the proposed duty of care of the White Paper for being too broadly framed (Graham, 2019; Nash, 2019; Tambini, 2019)[256]. If not correctly framed, such a duty of care may create a chilling effect on online speech (Tambini, 2019). It may also not be consistent with the prohibition of proactive measures under Article 15 of the ECD[257]. Another objection is the broadness and vagueness of the definition of the harms, as the White Paper aims at tackling a long list of harms, covering illegal but also harmful (legal) content (Nash, 2019; Tambini, 2019).

[256] Also Graham, A Ten Point Rule of Law Test for a Social Media Duty of Care, 2019, available at: https://www.cyberleagle.com/2019/03/a-ten-point-rule-of-law-test-for-social.html.

[257] Graham, Take care with that social media duty of care, 2018, available at: https://www.cyberleagle.com/2018/10/take-care-with-that-social-media-duty.html.

ANNEX II: LIST OF INTERVIEWED STAKEHOLDERS

The table below provides an overview of all the stakeholders that have been interviewed by the study team. In total, 24 stakeholders agreed to participate to the consultation and provided their input. This means that for the purpose of this study, 9 online platforms, 6 industry/trade associations, 1 hotline, and 8 NGOs have been interviewed.

Type of stakeholder	Organisation
Online platform	eBay
Online platform	Facebook
Online platform	Google
Online platform	JustPaste.it
Online platform	Microsoft
Online platform	Mozilla
Online platform	Olx
Online platform	Snap
Online platform	YouTube
Industry/Trade association	European Consumer Organisation (BEUC)
Industry/Trade association	Computer and Communications Industry Association (CCIA)
Industry/Trade association	European Internet Services Providers Associations (EuroISPA)
Industry/Trade association	News Media Europe
Industry/Trade association	Communication Agencies Association
Industry/Trade association	European Digital Media Association (EDiMA)
NGO	Center for Democracy & Technology (CDT)
NGO	Counter Extremism Project (CEP)
NGO	Cyber Data Coalition
NGO	European Digital Rights
NGO	Fundacja Panoptykon
NGO	Renaissance Numérique
NGO	Respect Zone Against Cyber Violence
NGO	Rettighedsalliancen
Hotline	Point de Contact

ANNEX III: QUESTIONNAIRE TO ONLINE PLATFORMS

Measures to moderate illegal content & effectiveness

1. How and what measures does your platform deploy to **distinguish legal from illegal content online**?
   - Are your **Terms of Service/Terms of Use** stricter than the legal rules, in identifying content to be removed?
   - Do you see an important **fragmentation** between the Member States on what illegal content is? If yes, for which content in particular? Does such fragmentation affect your operations and if so, how?
2. What are the **measures** you put in place to **detect and remove illegal content**?
   - How far are those measures **automated**? What are the pros and cons, the opportunities and the risks for Artificial Intelligence tools for content moderation?
   - Which **transparency** policies do you have?
   - Which safeguards did you put in place to protect **fundamental rights**?
   - Do you differentiate your measures according to the type of illegal content?
3. Did the policies put in place by your platform to moderate content **contribute to reducing the aggressiveness and quantity of illegal content**? Please explain how and provide relevant data/sources to support your reply.
4. What is **the impact of measures** you deployed in relation to illegal content moderation on:
   - making information accessible;
   - facilitating communication and interaction;
   - increasing choice of products and services; and
   - accessing new market and business opportunities? Please provide data.
5. How do you ensure that decisions to remove illegal content from your platform are **accurate and well-founded**, especially when automated tools are used?

Involvement of platforms' users

6. What **complaint mechanisms** have you implemented for users of your platform to report on illegal content? What is the share of content taken down as a result of the reporting by users?
7. What are the pros and cons, the opportunities and the risks of having content providers be able to give their views to your platform on the alleged illegality of the content through **a 'counter-notice' procedure**? Please give reasons for your answer and where available, quantitative indicators.
8. Once a decision is taken against illegal content such as depublication, delisting, downranking, or censorship of information or accounts; **how long does this measure remain effective**? How do **you address the stay-down issue**? Do you think EU law should address such issue beyond the recent Facebook case?
9. Do you **warn the users** of your platform when a content they flagged as illegal has been taken down or censored?

Challenges & potential solutions

10. What are the **challenges** faced by your platform to enforce legal rules and/or private regimes (e.g. Terms of Service/Terms of Use) on the moderation of illegal content online?
11. What do you think could be **done/implemented at the EU level** to improve the moderation by online platforms of illegal content?
    - In particular, do you think different rules should be adopted for **different types of illegal content**?
    - How should the increasing use of **automated tools** be addressed?
    - How should **fundamental rights** be protected and well balanced?
12. Do you estimate that the **different content moderation practices in EU Member States** may hinder or, to the contrary, facilitate the fight against illegal content in the EU overall? Please explain. Do you think more harmonisation should be imposed at the EU level?
13. In which **areas of the EU Internal Market** (e.g. the Digital Single Market; the area of freedom, security and justice; the free movement of services; the access to business opportunities and cultural richness) and its regulatory framework, do you consider **reforms necessary** to address existing or upcoming barriers or inefficiency/ineffectiveness of current legal solutions regarding online platforms' illegal content moderation?
14. Could you please provide **examples of best international practices** regarding online platforms' illegal content moderation which can serve as models to follow at European level?

Others

15. In the context of a limited exemption of certain categories of Internet intermediaries from secondary liability[258] are the **existing liability principles** of intermediary service providers (on which Section IV of the e-Commerce Directive 2000/31/EC is based) **fit-for-purpose** (Articles 12-15)? Please provide reasons for your answer.
16. What kind of content moderation mechanism should be implemented to respect the **freedom of expression and information**?
17. How to ensure the **fundamental right of not to be discriminated against** (prohibition of discrimination) is respected? How to hamper the development of new forms of online discrimination (e.g. homophobic remarks on online games' platforms)?
18. Is there **anything else you would like to add** in the context of online platforms' illegal content moderation practices?

[258] i.e. liability resulting from illegal users' behaviour, which has different possible rationales: promoting the activity of the intermediaries, preserving their business models, preventing excessive collateral censorship (i.e. preventing the intermediaries from censoring the expressions of their users).

ANNEX IV: QUESTIONNAIRE TO OTHER STAKEHOLDERS

Measures to moderate illegal content & effectiveness

1. Have you put in place any **voluntary or proactive measures to control** certain categories of illegal content from your system?
2. How **effective** do you estimate the measures deployed by online platforms have been in moderating illegal content? Please provide information and data/sources to justify your reply.
3. How effective do you consider the measures deployed by online platforms to **distinguish legal from illegal** content online? Please provide information and data/sources to justify your reply.
4. What is the **impact of measures** deployed in relation to online platforms' illegal content moderation on:
   - making information accessible;
   - facilitating communication and interaction;
   - increasing choice of products and services; and
   - accessing new market and business opportunities? Please provide data.

Involvement of platforms' users

5. Should online content providers be able to give their views to the hosting service on the alleged illegality of the content through a **'counter-notice'** procedure? Please provide reasons for your answer.
6. How **effective** are the **measures deployed by platforms to enable users** to report on illegal content?
7. Which types of measures should be taken to **improve the transparency** of platforms' decisions regarding illegal content reported by users?

Challenges & potential solutions

8. What are the **challenges** which you are faced with **in reporting** illegal content from an online platform?
9. What are the **challenges** which **online platforms** and other Internet intermediaries face in enforcing legal rules and/or private regimes (e.g. Terms of Service/Terms of Use) on illegal online content moderation?
10. What are the **possible solutions**, new measures to improve the EU regulatory framework and its enforcement regarding online platforms' illegal content moderation?
11. In which areas of the **EU Internal Market** (e.g. the Digital Single Market; the area of freedom, security and justice; the free movement of services; the access to business opportunities and cultural richness) and its regulatory framework, do you consider **reforms necessary** to address existing or upcoming barriers or inefficiency/ineffectiveness of current legal solutions regarding online platforms' illegal content moderation?
12. Is there a need to impose **a specific duty of care** regime on online platforms and other Internet intermediaries for certain categories of illegal content online?

    If yes, please specify:

- What are the **categories of illegal content** online requiring a specific duty of care?
- What **actions** (i.e. scope and format) shall constitute such a specific duty of care in relation to active and passive online intermediaries?
- Which measures shall be deployed to increase **transparency** related to the duties of care for online intermediaries with regard to the general content restrictions policies and practices by online intermediaries?

13. Do you estimate that the **different content moderation practices in EU Member States** may hinder or, to the contrary, facilitate the fight against illegal content in the EU overall? Please explain.
14. Could you please provide **best international practices** regarding online platforms' illegal content moderation which can serve as models to follow at European level?

Others

15. In the context of a limited exemption of certain categories of Internet intermediaries from secondary liability[259], are the **existing liability principles** of intermediary service providers (on which Section IV of the e-Commerce Directive 2000/31/EC is based) **fit-for-purpose** (Articles 12-15)? Please give reasons for your answer.
16. What kind of content moderation mechanism should be implemented to respect the **freedom of expression and information**?
17. How to ensure the **fundamental right of not to be discriminated against** (prohibition of discrimination) is respected? How to hamper the development of new forms of online discrimination (e.g. homophobic remarks on online games' platforms)?
18. Is there **anything else you would like to add** in the context of online platforms' illegal content moderation practices?

[259] i.e. liability resulting from illegal users' behaviour, which has different possible rationales: promoting the activity of the intermediaries, preserving their business models, preventing excessive collateral censorship (i.e. preventing the intermediaries from censoring the expressions of their users).

Online platforms have created content moderation systems, particularly in relation to tackling illegal content online. This study reviews and assesses the EU regulatory framework on content moderation and the practices by key online platforms. On that basis, it makes recommendations to improve the EU legal framework within the context of the forthcoming Digital Services Act.

This document was provided by the Policy Department for Economic, Scientific and Quality of Life Policies at the request of the committee on Internal Market and Consumer Protection (IMCO).

PE 652.718
IP/A/IMCO/2019-10

Print ISBN 978-92-846-6793-2 | doi:10.2861/194019 | QA-04-20-302-EN-C
PDF ISBN 978-92-846-6792-5 | doi:10.2861/831734 | QA-04-20-302-EN-N