TASK FORCE REPORT

TOWARDS A PRINCIPLED LEVEL PLAYING FIELD FOR AN OPEN AND SECURE ONLINE ENVIRONMENT

Regulation, enforcement and oversight of online content moderation in the EU and the United Kingdom

October 2022

SERGIO CARRERA
VALSAMIS MITSILEGAS
MARCO STEFAN
NIOVI VAVOULA

SUMMARY

This Task Force Report examines the regulation, oversight, and enforcement of online content moderation in the European Union and the United Kingdom. It identifies the key issues and challenges related to co-regulation and self-regulation of content moderation standards, including internal oversight mechanisms of online platform services. The Report examines regulatory content moderation standards that exist internationally, in the EU, and the UK. It assesses the key issues and open questions characterising the roles and responsibilities carried out by independent oversight and regulatory authorities, and the main challenges that online content moderation policies and practices raise to fundamental rights, democracy, and rule of law. Particular attention is given to assessing issues related not only to privacy and data protection, but also to those affecting freedom of expression and the rule of law, due process, and effective remedies. The Report advances a set of policy recommendations aimed at ensuring a principled level playing field for an open and secure online environment.

Sergio Carrera is a Senior Research Fellow and Head of Unit at the Centre for European Policy Studies (CEPS), Valsamis Mitsilegas is a Professor at the School of Law and Social Justice in the University of Liverpool (UK), Marco Stefan is a former Research Fellow at CEPS, and Niovi Vavoula is a Lecturer at the School of Law of Queen Mary University London (QMUL).

This Report presents the key findings and policy recommendations emerging from a Task Force project funded by the UK Mission to the EU. The views expressed in this Report are those of the authors and do not necessarily reflect those of the UK Mission to the EU or any institution to which the authors are associated.

CEPS • Place du Congrès 1 • B-1000 Brussels • Tel: (32.2) 229.39.11

TABLE OF CONTENTS

EXECUTIVE SUMMARY

A current source of international regulatory and law enforcement concern is to prevent online and social media platforms from being used as vectors for the production, exchange, and dissemination of illegal and harmful content online. Misuses of the internet for criminal purposes have progressively highlighted the ever-expanding societal and legal responsibilities of these platforms. Technological developments and the constant increase in the volume of personal data shared online have led these service or platform providers to acquire an increasingly prominent role in pre-empting and tackling crime online.

Private sector's self-regulatory efforts related to the moderation and removal of online content have progressively led to the development and implementation of a wide range of digital surveillance policies and practices. At the same time, online platform providers have been facing unprecedented pressure to comply with content regulation demands emanating from policies and norms operating at different levels of governance, thus creating various forms of 'public-private partnerships'.

A number of legislative initiatives have emerged in the European Union (EU) to deal with the proliferation of illegal content including child sexual abuse material, hate speech, commercial scams and frauds, breaches of intellectual property rights, and terrorist content online. These include for instance the Proposal for a Regulation on the Removal of Terrorist Content Online (TERREG), the Digital Services Act (DSA), and more recently the Child Sexual Abuse Material (CSAM) proposal. In the United Kingdom (UK), the Online Safety Bill aims at defining what types of 'illegal' as well as of 'legal but harmful' content platforms will have to tackle. The Bill introduces specific online content moderation and removal obligations for online platforms providers, and related sanctions for non-compliance.

Norms, policies, and practices that justify the use of personal data and new technologies by online platforms providers at various governance levels and across jurisdictions for the purpose of fighting crime and 'preventing harm' in the online environment have profound implications on the triangular relationship between fundamental rights— including the right to private life, data protection and the freedom of expression— as well as the rule of law and democracy.

Furthermore, independent oversight and regulatory bodies remain ultimately responsible for guaranteeing fundamental rights, including through the delivery of effective— administrative and judicial— remedies in case of abuses. With the proliferation of online surveillance duties and content moderation practices by a wide range of private and public actors, serious operational challenges emerge regarding the ability of competent oversight and regulatory authorities to effectively monitor compliance with existing standards and provide access to redress to individuals affected by wrongdoings.

This report is based on the Task Force that was formed between January and April 2022 by the Justice and Home Affairs (JHA) Section of the Centre for European Policy Studies (CEPS), in partnership with the Global Policy Institutes at Queen Mary University of London GPI-QMUL),

the University of Liverpool and the RENFORCE Centre at Utrecht University. The Report reflects the key issues and findings that emerged from Task Force meetings and deliberations.

Section I deals with the key issues and challenges related to co-regulation and self-regulation of content moderation standards, including internal oversight mechanisms. The following key findings have been highlighted:

- Online platforms' (co or self)-regulation of online content moderation, entails the identification of what is to be considered as prohibited content or disallowed speech. Such identification depends on what specific laws oblige social networks and online platforms to consider as 'manifestly unlawful' or 'harmful'. However, there exists an inherent legal imprecision and a lack of clarity characterising some of the key definitions, which tend to be too broad or over-inclusive. Furthermore, there exists the risk that online platforms providers become vehicles or proxies of both over-implementation and over-enforcement as a result of legal uncertainty as to the exact scope and content of their obligations due to the multiplicity of applicable rules, which may also hinder their effectiveness.
- Besides, internal rules such as those provided by Companies' Terms of Use, Terms of Services, or Terms and Conditions (ToS) or community guidelines, have been considered to be too vague and broad. There is a clear trend across the largest platforms towards greater complexity in their content moderation policies. Both ToS and community guidelines are regularly subject to changes in terms of content moderation policies making it difficult for users to properly understand and adhere to company policies. This also effectively hinders regulators, oversight bodies, and civil society actors' capacity to hold the platforms accountable for their enforcement decisions. Finally, the lack of publicity contributes to the opacity and lack of accountability and transparency.
- Platforms combine human moderation with automated decision-making processes through the deployment of various Artificial Intelligence (AI) systems. A hybrid approach is often preferred, which typically involves the automatic identification or pre-filtering of content that is then checked by a team of human moderators. This notwithstanding the fact that the quality and reliability of content moderation performed by AI tools remains limited in many respects, the reproduction and amplification of biases and the fact that there is often no way of meaningfully knowing whether some discriminatory results occur as a result of AI operations in the online content moderation domain.
- Content moderation is also developing through multilateral fora aimed at increasing private sector coordination. However, concerns have been voiced about the structural lack of external oversight of industry cooperation in online content policing and the lack of transparency related to scope, scale, and impact of such practices. There is also an absence of robust, transparent, and reliable mechanisms for due process available to users.

- Online content moderation is carried out without independent judicial or administrative oversight, but some platforms have established with internal mechanisms to complaint against decisions made about users' content. The example of the Facebook Oversight Board is paradigmatic in this regard. Though such efforts are encouraged at international level, concerns have been voiced about the Board's legitimacy as a non-state mechanism linked to the protection of freedom of expression.

Section II provides a detailed examination of existing regulatory content moderation standards that exist internationally, in the EU, and the UK. The Section begins with the international level, where a number of initiatives and calls have emerged in the past few years, such as the work conducted by the UN Rapporteur on the promotion and protection of the right to freedom of opinion and expression, the OECD and the Recommendation by the Council of Europe. The Section then explores the EU legal framework from its first modest steps to coordinate national actors; in particular the EU Internet Forum and EU Internet Referral Unit, established within Europol, both developed in the context of EU counter-terrorism efforts. The main findings are the following:

- TERREG, adopted in 2021, obliges online service platforms to act upon targeted pieces of online content, in striking resemblance to judicial practice. Service providers have proactive duties in cases of 'exposure' to terrorist content and on the basis of their own terms and conditions. The removal of terrorist content, or disabling access to it following a request, must be done as soon as possible and in any event within one hour of receipt of the removal order. This very tight deadline, which may incentivise the providers to use automated content moderation tools in order to identify and delete terrorist content, has been a key criticism.
- Another EU legal instrument of relevance is the amended Europol Regulation, which expands the scope of law enforcement authorities cooperation with service providers in the fight against crime and terrorism, including in the field of digital surveillance of online content. Europol will be able to receive personal data directly from private parties and analyse that data to identify those Member States that could open investigations into related crimes. The agency can support Member States' actions to address dissemination of terrorist content in the context of an 'online crisis' situation, such as that stemming from online dissemination of child sexual abuse material.
- The latest initiative at EU level is the Commission-proposed Regulation on Child Sexual Abuse Material (CSAM), according to which providers will be obliged to detect, report, and remove or disable child sexual abuse material on their services. This will not only concern child sexual abuse material which has been verified as such by authorities. Providers must additionally proactively search for new photos and videos, as well as evidence of text-based 'grooming' which will require use of AI-based tools and techniques to scan private conversations. The proposal has been condemned for embracing scanning and surveillance technologies for identifying illegal content and essentially granting the possibility to access content of communications on generalized and indiscriminate basis and seriously interfering with the essence of the rights of privacy and data protection.

- Perhaps the most important EU initiative concerns the recently agreed Digital Services Act (DSA), which amends the e-Commerce Directive and lays down EU-wide due diligence obligations that will apply to all digital services that connect consumers to goods, services, or content depending on their roles, size, and impact on the online ecosystem. The DSA has maintained the general rule of the e-Commerce Directive according to which providers of hosting services are not liable for user-generated content unless they have actual knowledge about illegal online activity or upon obtaining such knowledge or awareness, act expeditiously to remove or to disable access to the illegal content.
- A key component of these duties concerns new procedures for fast removal of illegal content. Providers must act against illegal content to comply with national orders and provide information about the action taken. In addition to these requirements, the DSA imposes new mechanisms allowing users, both persons and legal entities, to flag illegal content online and access to an effective internal complaint-handling system under specific rules for the processing of complaints. Supervision of very large companies is bestowed to the Commission. Member States are required to set up Digital Services Coordinators, which will have investigatory powers to require providers to provide information relating to infringement of the DSA. An independent group composed of Digital Services Coordinators and chaired by the Commission, the European Board for Digital Services, is tasked with supervising providers.
- With regard to the UK, an Online Safety Bill is currently under negotiation. The Online Safety Bill pushes for more obligations on so-called 'regulated services'—that is 'user-to-user services' and 'search services' that have 'links' with the UK— with regard to three types of content: (a) illegal content; (b) content that is harmful to children; and (c) content that is legal but harmful to adults. As with the DSA, the specific scope of the duties varies significantly depending on the nature of the service and the nature of the content. The Office of Communications (Ofcom) will be entrusted with enforcement tasks and must prepare codes of practice to assist service providers comply with their duties.

Section III of this Report covers the key issues characterising the roles and responsibilities carried out by independent oversight and regulatory bodies in the EU and the UK. The most relevant findings can be summarised as follows:

- A key characteristic of regulatory oversight and accountability bodies is that of being independent from both regulators and online platforms providers. Given that the implementation of online content moderation activities entails compliance with a multi-level normative and policy framework a mosaic of bodies currently comprise the framework of regulatory oversight that apply to online content moderation.
- National Data Protection Authorities (DPAs), through their investigative and corrective powers in the application of EU data protection law, provide a crucial network of regulatory oversight and external accountability framework covering data processing (regardless of whether it is voluntary, or legally mandated) in online content moderation. However, there also are clear limitations to the DPAs capacity to effectively oversee online platforms activities, and to prevent or redress the different abuses.

- An intrinsic limitation derives from the very nature and scope of the DPAs mandate, which exclusively relates to the effective and consistent application of EU or national laws in the field of data protection. Another critical limitation derives from the lack of adequate resources or capacity to ensure effective monitoring of correct implementation of the General Data Protection Regulation (GDPR) and/or the Law Enforcement Directive (LED). Finally, pieces of EU legislation, such as TERREG (which regulates various forms of private-public cooperation in the online content moderation domain), do not expressly foresee a role for DPAs.
- The requirement in the DSA to set up DSCs is a step forward. Yet, the DSA pays little explicit attention to other relevant enforcement authorities to address situations of potential overlap in competencies and lacks institutionalised and structured cooperation between other competent oversight authorities in matters of mutual concern. One common concern in this regard is that the DSA fails to provide a clear legal basis for the DSCs, the European Board for Digital Services (EBDS), and the Commission to cooperate and/or consult with other EU enforcement networks. The involvement of other regulatory bodies is generally discretionary: the DSA leaves it at the Member State discretion to provide for regular exchanges of views with other authorities.
- As for the Online Safety Bill, which designates Ofcom as the regulatory authority entrusted with oversight, Ofcom provides a model for a regulatory agency, which is targeted at communications and has a wider mandate that includes but goes beyond the rights of privacy. However, it will have a difficult balancing act to manage the different interests at stake, and to be an effective but not overly heavy-handed regulator. Besides, in its current form, the Online Safety Bill makes Ofcom, which is an independent regulator, overly dependent upon the Secretary of State, which undermines its status and will affect its regulatory oversight and the overall performance of its duties.

Section IV of the Report identifies and examines the most crucial fundamental rights, democracy and rule of law issues emerging from online content moderation policies and practices in relation to privacy and data protection, freedom of expression and the rule of law, due process and effective remedies, which can be synthesized as follows:

- With regard to privacy and data protection, legislative developments, such as TERREG and CSAM, as well as the Online Safety Bill, have raised a challenge of *generalised monitoring of online content.* EU Member States have also introduced measures that impose or enable large-scale online surveillance by private platforms, such as the German NetzDG. Companies' development of capabilities for proactively screening content could also limit platforms' ability to encrypt private messages, disrupting their business model and removing their competitive edge in their respective market. In light of the Court of Justice of the EU (CJEU) jurisprudence, effective safeguards are necessary in the legal framework to ensure that there is an appropriate framing of the situation to ensure that monitoring of online content/certain systemic risks— including through the use of automated data processing— takes place in conformity with applicable privacy and data protection rules. For instance, this would mean the inclusion of provisions qualifying the

types of illegal content that may warrant use of automated detection techniques involving the processing of personal data and delineating the circumstances in which voluntary notification may take place.

- Furthermore, public authorities' engagement may take the form of direct surveillance of information on online platforms through the probing of 'risk individual' or 'risk communities'. As a result, the Report underlines the need to ensure an *ex ante* independent scrutiny of law enforcement access to data for criminal investigation-related purposes.
- Another specific data protection-related challenge arises from the fact that online content moderation often requires the processing of a wide range of personal information resorting to Ai-based tools. To be in line with the GDPR, particularly its Article 22 on automated decision-making, the use of AI-based tools must be accompanied by human oversight and verification mechanisms as a safeguard that the decisions taken are accurate and well-founded. In that respect, the increased transparency requirements in the DSA are a positive development.
- Freedom of expression concerns are raised in relation to the scope of the illegal content in online moderation, which often, fails to provide an exact or sufficiently precise definition of the type of content that actually qualifies as 'illegal'. The key risk in such cases is that the law serves as tool of online censorship also through criminalisation directed at shaping the online regulatory environment and at suppressing legitimate discourse. For example, the broad definition of illegal content entails that the DSA does not impose any limits as to what content can be criminalised at the national level and the clarifications do not help in this regard. In turn, in the case of the UK, the listed offences seem to be targeted and limited to what is strictly necessary and framed under clear terms, thus the UK approach appears more balanced than the one outlined in the DSA.
- Another source of freedom of expression-related concern involves the moderation of online content considered to be "harmful" but not illegal. Serious legal certainty challenges arise in a context where there is a lack of common understanding of what exactly constitutes harmful (but not illegal) online content. As a way of illustration, the DSA doesn't define what content is 'harmful', whereas the definition provided in the Online Safety Bill as to what constitutes 'content that is harmful to adults' does not meet the legality requirement under international human rights law. It is too vague and unclear and does not enable individuals to foresee what content will be moderated and therefore what they can or cannot write online, particularly to individuals they do not know.
- As for the use of automated means, there is need for strict human review due to the limitations and errors of AI technology. Otherwise, the increasing reliance on AI-based tools will entail that a number of erroneously flagged content will be removed, infringing freedom of expression.

- As for risks to due process, rule of law, and effective remedies, the requirement under the TERREG Regulation to remove allegedly illegal content in a particularly short time frame does not allow for the possibility to turn to the court and empowers platforms with extra power without proper oversight. The issue of judicial review of decisions on removal, and thus of effective oversight, has been left unanswered. Besides, a very large degree of discretion is still left to companies with regard to the level of independence, accessibility, transparency, and predictability of their in-house appeal mechanisms in the context of internal oversight mechanisms and review procedures.

Finally, *Section V* of the Report concludes and puts forward a set of policy recommendations along three main lines: First, ensuring regulatory clarity and quality, and upholding the principle of legality; Second, ensuring accessible, foreseeable, and transparent online platform services' monitoring and oversight policies; and third, guaranteeing effective regulatory and networked oversight, and enhancing multi-actor coordination.

ABBREVIATIONS

AADC	Age Appropriate Design Code
ACHPR	African Commission on Human and Peoples' Rights
AI	Artificial Intelligence
AVMSD	Audio-visual Media Service Directive
CJEU	Court of Justice of the EU
CMA	Competition and Markets Authority
CoE	Council of Europe
COM	European Commission
CSAM	EU Regulation on Child Sexual Abuse Material
CSAM	Child Sexual Abuse Material
CSEP	Civil Society Empowerment Programme
CTIRU	Counter Terrorism IRU
DMA	Digital Markets Act
DPAs	National Data Protection Authorities
DRCF	Digital Regulation Cooperation Forum
DSA	Digital Services Act
DSC	Digital Services Coordinator
ECHR	European Convention of Human Rights
ECtHR	European Court of Human Rights
ECTC	Europol's European Counter Terrorism Centre
EDBS	European Board for Digital Services
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EMSC	Europol's European Migrant Smuggling Centre
Europol	European Agency for Law Enforcement Cooperation

FCA	Financial Conduct Authority
FEOs	Freedom of Expression Officers
FOB	Facebook Oversight Board
GDPR	EU General Data Protection Regulation
GIFCT	Global Internet Forum to Counter Terrorism
ICCPR	International Covenant on Civil and Political Rights
ICO	Information Commissioner's Office
IMCO	European Parliament's Internal Market and Consumer Protection Committee
IRUs	Internet Referral Units
JHA	Justice and Home Affairs
JURI	European Parliament Legal Affairs Committee
LED	Law Enforcement (Data Protection) Directive
LGBTQ+	Lesbian, gay, bisexual, transgender, queer/questioning and others
LIBE	European Parliament Civil Liberties, Justice and Home Affairs Committee
OAS	Organisation of American States
OECD	Organisation for Economic Cooperation and Development
Ofcom	UK Office of Communications
OHCHR	United Nations Office of the High Commissioner for Human Rights
OJ	Official Journal of the European Union
OSCE	Organization for Security and Co-operation in Europe
TERREG	EU Regulation on the Removal of Terrorist Content Online
ToS	Terms of Services / Terms and Conditions
VLOPs	Very Large Online Platforms
VSPs	Video-Sharing Platforms

INTRODUCTION
SETTING THE SCENE AND THE SCOPE OF THE TASK FORCE

In the digital age, private and public interactions are largely enabled by products and services developed and made available by providers of Information and Communication Technologies (ICTs). Online platforms and social media enable all kinds of social and economic relationships and constantly provide **new spaces for open public discourse, civic participation, and democratic engagement and mobilisation**.

At the same time, the 'datafication' of contemporary societies raises crucial legal and ethical challenges. Together with the constant growth of online interactions, the increasing uses of digital communication tools (and of the volumes of data produced through them)— expose data subjects— in their quality of citizens and consumers alike— to new risks. The digital revolution has generated great possibilities for misuses and abuses of ICTs in ways which can profoundly affect democratic societies as well as individual freedoms and rights.

A current source of international regulatory and law enforcement concern is to prevent online and social media platforms from being used as vectors for the production, exchange, and dissemination of illegal and harmful content online. Misuses of the internet for criminal purposes have progressively highlighted the ever-expanding societal and legal responsibilities of these platforms. Technological developments and the constant increase in the volume of personal data shared online have led these service or platform providers to acquire a more prominent role in pre-empting and tackling crime online[1].

To improve safety on their platforms and across their services, tech companies have put in place their own codes of conduct and/or community guidelines. **Private sector's self-regulatory efforts related to the moderation and removal of online content have progressively led to the development and implementation of a wide range of digital surveillance policies and practices**[2]. Work related to the tackling of illegal content has also been developed in the context of various multilateral fora launched by representatives of the internet industry[3].

Online platform providers have been facing unprecedented pressure to comply with content regulation demands emanating from policy and norms makers operating at different levels of governance. Service providers' own undertakings to address the posting and dissemination of illegal and harmful content have advanced in parallel with a wide range of different policy and normative measures at the national, regional, and international level. Under these measures, **the detection and removal of illegal and harmful content online relies on the establishment and**

[1] For an in-depth investigation into media platforms' role in policing of online content, see Gillespie, T. (2018), *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*, Yale University Press; Huszti-Orban, K. (2017), *Countering Terrorism and Violent Extremism Online: What Role for Social Media Platforms?*, Rio de Janeiro, Fundação Getulio Varga.

[2] Roberts, S. T. (2019), *Behind the screen: Content moderation in the shadows of social media*, Yale University Press.

[3] For instance, the Global Internet Forum to Counter Terrorism (GIFCT), created by Facebook, Google, Twitter, and Microsoft as part of a commitment to increase industry collaboration to combat illegal online hate speech.

implementation of various types of 'public-private partnerships'. Public-private cooperation has been at the centre of EU attempts to fight the proliferation of illegal content — i.e., information which is not in compliance with EU law or the law of a Member State, including child sexual abuse material, hate speech, commercial scams and frauds, breaches of intellectual property rights, and terrorist content online.

Since 2015, the European Union Agency for Law Enforcement Cooperation (Europol) disposes of Internet Referral Units (IRUs) that are responsible for referring 'terrorist and violent extremist content' to online service providers[4]. In 2017, the Commission formulated a set of political guidelines on tackling illegal content online[5], and in 2018 it issued a Recommendation indicating the 'operational measures' that should be taken by companies and Member States regarding the detection and removal of illegal content through reactive (so-called 'notice and action') or proactive measures[6]. Actions concerning social media platforms and online content are currently part of the 'core vision for European security'[7], featuring in both the 2020 Security Union Strategy[8], and the EU Counter-Terrorism Agenda[9].

The adoption of EU Regulation on the Removal of Terrorist Content Online[10], also known as TERREG, constitutes a key development in the establishment of an EU normative framework on private-public cooperation in the field of online content moderation and removal of content online. The Regulation assigns service providers with 'law enforcement duties' to remove, disable access to, or assess nature of online content in ways that are both reactive (i.e., centred around compliance with competent authorities' orders[11]) and proactive (i.e., based on provisions included in the provider's own terms and conditions[12]). In order to enable the detection and identification and removal of content by online service providers, TERREG also legitimises automation in content moderation and the use of machine learning tools to perform online surveillance duties[13].

---

[4] Between July 2015 and 2018, Europol made over 50 000 decisions on referrals to service providers about terrorist content on their platforms. To mention just one example, the UK's Internet Referral Unit (CTIRU) alone identified 300 000 pieces of terrorist content between 2010 and 2018. European Commission, Impact assessment accompanying the proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online, SWD(2018) 408, 12 September 2018.

[5] European Commission, Communication, 'Tackling Illegal Content Online. Towards an enhanced responsibility for online platforms, COM(2017) 555, 28 September 2017.

[6] European Commission, Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, OJ [2018] L 63/50.

[7] Bellanova, R. and de Goede, M. (2021), 'Co-Producing Security: Platform Content Moderation and European Security Integration', *Journal of Common Market Law Studies*, p. 2.

[8] European Commission, Communication on the EU Security Union Strategy, COM(2020) 605, 24 July 2020, p. 13-14.

[9] European Commission, Communication, A Counter-Terrorist Agenda for the EU: Anticipate, Prevent, Protect, Respond, COM(2020) 795, 9 December 2020.

[10] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online.

[11] Ibid. Art. 4.

[12] Ibid. Art. 5.

[13] Ibid. Recital 25.

New EU measures to counter illegal products, services and content online, including procedures for removal by the private sectors will be introduced by the Digital Services Act (DSA)[14]. The DSA, which amends existing EU horizontal legislation[15] and complements sectoral legislation[16] regulating the production and exchange of content online, envisages *inter alia* that 'very large online platforms' (VLOPs) will be subject to specific obligations due to the 'particular risks they pose in the dissemination of both illegal and harmful content'[17]. In parallel, the recast of the Europol Regulation[18] expands the scope of law enforcement authorities cooperation with service providers in the fight against crime and terrorisms, including in the field of digital surveillance of online content[19].

Several non-EU countries have or are in the process of adopting new legislation on tackling illegal as well as harmful content online. In the United Kingdom (UK), the Online Safety Bill[20] aims at defining what types of 'illegal' as well as of 'legal but harmful' content platforms will have to tackle. The Bill also envisages the introduction of specific online content moderation and removal obligations for service providers, and related sanctions for non-compliance.

The objective to control, prevent, and pre-empt the production and dissemination of illegal or harmful content online involves **the collection by the private sector and subsequent processing by both private and public actors on a large scale** (and in particular through the automated processing of data by machine-learning technologies) **of personal data emanating from everyday communications and interactions**. The increasing attention that is paid at national and supranational fora to the monitoring of online content production and dissemination clearly shows how, in a world of large-scale data accumulation in the digital spheres, there is a 'confluence of surveillance interests between the public and the private sector'[21].

---

[14] European Commission, Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020) 825, 15 December 2020.

[15] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), OJ [2000] L178/1.

[16] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audio-visual media services (Audio-visual Media Services Directive) in view of changing market realities OJ [2018] L303/69; Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ [2019] L130/92.

[17] European Commission, Proposal for DSA Act, Recital 63.

[18] European Commission, Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) 2016/794, as regards Europol's cooperation with private parties, the processing of personal data by Europol in support of criminal investigations, and Europol's role on research and innovation, COM(2020) 796, 9 December 2020.

[19] Ibid. Article 1, Article 18a.

[20] A Bill to make provision for and in connection with the regulation by OFCOM of certain internet services; for and in connection with communications offences; and for connected purposes (Bill 004 2022-23).

[21] Mitsilegas, V. (2021), 'The Privatisation of Surveillance in the Digital Age' in V. Mitsilegas and N. Vavoula (eds.), *Surveillance and Privacy in the Digital Age: European, Transatlantic and Global Perspectives*, Hart, pp. 101-158.

Measures adopted at the national, regional, and international fora present their own specificities in terms of the approach followed in the definition of ***what*** constitutes illegal or harmful content, the monitoring duties and enforcement roles assigned to service providers and law enforcement authorities, the roles and responsibilities of oversight authorities or regulators, as well as the tools, mechanisms, and procedures through which certain content could be identified, flagged, and removed. With the emergence of a multi-level and multi-actor playing field related to surveillance, moderation and removal of online content, crucial questions of legal certainty and coherence, as well as effective protection of fundamental rights of individuals arise.

Norms, policies and practices that justify the use of personal data and new technologies by private and public actors operating at various governance levels and across jurisdictions for the purpose of fighting crime and 'preventing harm' in the online environment have profound implications on the triangular relationship between fundamental rights, the rule of law and democracy[22]. The three are understood as *co-constitutive* and as a trinity reflected in a triangular interaction which ensures the legally based rule of a democratic State that delivers fundamental rights[23].

Data-driven moderation and removal of content online generate tensions and interference with the rights to respect for private life (Article 7 of the EU Charter of Fundamental Rights) and data protection (Article 8 EU Charter), which are intimately linked to human dignity in the digital age. Moderation of online content also has serious and potentially chilling effects on the freedoms of expression and information of internet users (Article 11 EU Charter). It can negatively impact their freedoms of assembly and association (Article 12 EU Charter) and run against the prohibition of discrimination (Article 21 EU Charter). Online content moderation poses challenges to the safeguarding of the internet as a critical public space and as open market to European business (Article 16 EU Charter). Some of these rights have been granted further normative substance through the enactment of EU secondary legislation such as for instance the EU General Data Protection Regulation (GDPR)[24].

[22] Carrera, S., Guild, E. and Hernanz, N. (2013), 'The Triangular Relationship between Fundamental Rights, Democracy and the Rule of Law in the EU: Towards an EU Copenhagen Mechanism', Study for the European Parliament, Brussels.

[23] According to the Council of Europe Venice Commission's Rule of Law Checklist, 'The Rule of Law promotes democracy by establishing accountability of those wielding public power and by safeguarding human rights, which protect minorities against arbitrary majority rules.' European Commission for Democracy through Law (Venice Commission), Rule of Law Checklist, 18 March 2016.

[24] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ [2016] 119/1; See also Directive (EU) 2016/680 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 by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA OJ [2016] 119/89.

In addition, human rights, democratic and rule of law standards applying to the prevention of 'online crimes and harms' are also provided under legally binding human rights instruments and standards established at both **international (e.g., United Nations)**[25] **and regional (e.g., Council of Europe) venues**. And while rights such as data protection and freedom of expression are not absolute in nature and may allow for strictly defined derogations, in order **to prevent arbitrariness**, state authorities must justify that these are prescribed by **clear/precise, accessible and predictable/foreseeable—high quality—laws (legality principle), pursue a legitimate purpose and be necessary in a democratic society and proportionate**.

Against this backdrop, it is crucial to ensure that individuals have **'the same rights online as offline'**[26], and that the **legal safeguards anchored in national, regional, and international law applying to online content moderation and removal practices are effectively and consistently protected and enforced**.

In a context where an increasing number of actors acquire powers and responsibilities to control, prevent, and pre-empt the production and dissemination of terrorist, illegal, or harmful content online (in particular through the processing of data by machine-learning technologies), the roles of law enforcement authorities and tech companies in preventing and combatting the production and dissemination of terrorist and violent extremist content online must not only be clearly and consistently regulated, but also be monitored by the independent oversight bodies, both administrative and judicial in nature.

In democratic societies governed by the rule of law, **independent oversight and regulatory bodies remain ultimately responsible for guaranteeing fundamental rights, including through the delivery of effective— administrative and judicial— remedies in case of abuses**. However, the multiplication of online surveillance duties and content moderation practices by a wide range of private and public actors poses **serious operational and challenges to the ability that competent oversight and regulatory authorities have to effectively monitor compliance with existing standards and provide access to redress to individuals affected by wrongdoings**.

Scope, methodology, and objectives of the Task Force

Between January and April 2022, the Justice and Home Affairs (JHA) Section of the Centre for European Policy Studies (CEPS), in partnership with the Global Policy Institutes at Queen Mary University of London (GPI-QMUL), the University of Liverpool, and the RENFORCE Centre at Utrecht University set up a forum for a structured multi-stakeholder expert dialogue— Task Force— which:

[25] See for instance Article 18 of the International Covenant on Civil and Political Rights (ICCPR) which protects against arbitrary interference with a person's privacy. Refer also to Articles 19, 21, 22, 25 and 27 ICCPR which respectively guarantee the right to freedom of opinion and expression, peaceful assembly, freedom of association, taking part in public affairs and the rights of minorities.

[26] UN Human Rights Office Press briefing: Online content moderation and internet shutdowns, 14 July 2021, Press_briefing_140721.pdf (ohchr.org).

1. Discussed the different ways in which service providers currently perform content moderation tasks and enforce online safety-related duties, and highlighted the fundamental rights challenges raised by the performance of these tasks by private actors, including in cooperation with public authorities in the EU and the UK.
2. Analysed and compared ongoing policy and normative developments at the EU and UK in the field of online content surveillance and examined the ways in which these initiatives propose to frame and regulate public-private cooperation when addressing illegal and harmful content.
3. Discussed the organisational, and operational challenges that regulatory bodies and oversight actors operating in the EU and the UK respectively face when exercising control over the multitude of online harm surveillance duties that tech companies currently perform or may be required to undertake in future.
4. Assessed the extent to which policies and norms related to online harm and illegal content online can effectively guarantee the protection of fundamental rights and freedoms protected at the national, European, and international level.

The Task Force discussions involved representatives of the private sector, national and supranational regulatory and supervisory authorities, as well as academics and civil society actors (Task Force Members). The Task Force provided a neutral platform for independent legal and policy-research exchange and the collection of expert stakeholders' knowledge. This Report reflects the key issues and findings that emerged from the Task Force meetings and deliberations that took place on 23 February 2022 and 31 March 2022. This has been complemented by data collection and research by CEPS and GPI-QMUL Researchers. The Task Force members submitted comments on an earlier draft of this Report. Its contents reflect the general tone and direction of the discussion. The Reports findings and recommendations do not necessarily represent a full common position or consensus among all Task Force members or the views of any individual participant. The authors are solely responsible for the content, conclusions, and recommendations included in the Report. A full list of Task Force members and participants appears in Annex 1 of this Report.

The Report is divided into four main Sections: *Section I* deals with the key issues and challenges related to co-regulation and self-regulation of content moderation standards, including internal oversight mechanisms. *Section II* provides a detailed examination of existing regulatory content moderation standards that exist internationally, in the EU, and in the UK. *Section III* covers the key issues characterising the roles and responsibilities carried out by independent oversight and regulatory bodies in the EU and the UK. *Section IV* identifies and examines the most crucial fundamental rights, democracy and rule of law issues emerging from online content moderation policies and practices. *Section V* concludes and puts forward a set of policy recommendations.

SECTION I. CONTENT MODERATION BY ONLINE PLATFORMS: CO-REGULATION, SELF-REGULATION, AND RELATED CHALLENGES

'Content moderation' is used as a means to manage users' activities online[27]. It has increasingly become a tool to ensure 'online safety' by the means of **labelling, identifying, reporting and eventually deleting or removing content** defined as 'illegal' or 'terrorist' pursuant to national, international, or supranational laws and policies applicable in the— often multiple— jurisdictions under which online platforms operate. It is also used to tackle other forms of content considered to be not illegal but 'harmful' which may include hate speech, graphically violent or otherwise objectionable content, as well as various forms of disinformation[28]. **Content moderation by online platforms can therefore be instrumental to the pursuit of various policy goals**, ranging from the implementation of anti-terrorism and criminal justice policies to, *inter alia*, the enforcement of copyright laws and intellectual property rights.

Decision-making by social platforms in relation to moderation of illegal, terrorist, or other forms of content considered to be harmful may directly result from pressure (exerted through legal or policy means) from states or competent regulators operating at the national, international, supranational level. The development by private actors of online content moderation practices in response to competent public authorities' demand to address a public policy objective has been referred to as '**co-regulation**'[29]. Co-regulation encompasses situations where service providers (or associations thereof) adopt internal rules on content moderation **to comply with or adapt to specific and express legal obligations** to combat the dissemination of unlawful, terrorist, or harmful content online.

Regulation of online content is a clear trend worldwide[30]. At the national level, one such example of co-regulation is provided by the Network Enforcement Act (NetzDG) adopted by Germany in 2018[31]. The Act holds social media platforms accountable for combating speech deemed illegal under German law. It requires social media providers to take down content considered to constitute hate speech, or to support terrorism. Under the Act, systemic failure to remove content within the designated period is punishable by fines of up to 50 million

[27] Council of Europe (2021), Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation, Guidance Note Adopted by the Steering Committee for Media and Information Society (CDMSI) at its 19th plenary meeting, 19-21 May 2021, p. 3.

[28] In relation to regulating online content, the United Nations Office of the High Commissioner for Human Rights (OHCHR) has noted: 'While initial debates focused on copyright protection and online child sexual abuse, the discussion has now increasingly shifted to how to prevent the spreading of extremist content, hate speech and disinformation'. OHCHR (2020), 'Explainer: Regulating Content Online – the way forwards', November 2020.

[29] Marsden, C. T. (2012), 'Internet Co-Regulation and Constitutionalism: Towards European Judicial Review', *International Review of Law Computers & Technology*, Vol. 26, No 2, pp. 211-228.

[30] The OHCHR recently reported that some 40 new social media laws have been adopted worldwide in the last two years, and that another 30 are under consideration. See, OHCHR (2021), 'Moderating online content: fighting harm or silencing dissent?', 23 July 2021.

[31] Act to Improve Enforcement of the Law in Social Networks (2017).

euros[32]. The German framework for regulating online content constitutes one of the most extensive regulations of online content in the world. The law, which has paved the way for other countries to adopt similar forms of 'intermediary liability' legislation[33], has been widely criticised as vague and overinclusive as to the scope of material to be removed, and inadequate in terms of remedies provided and complaints mechanisms established[34].

In certain other EU countries, **the adoption of national laws entrusting service provides with obligations in the fight against (potentially) illegal online content has already raised constitutional issues reaching courts**. In France, most notably, the provisions of the so-called 'Avia law'[35] concerning the obligation imposed upon online service providers to remove content flagged by users as 'manifestly illegal' have been declared unconstitutional by the French Constitutional Council[36]. The latter *inter alia* held that several provisions of the law restricted the exercise of the freedom of expression in a manner that is not necessary, appropriate and proportionate[37].

In addition to the shaping of obligations related to online content moderation through hard law, co-regulation also includes **cases where online platforms and other internet intermediaries implement content moderation practices based on public policy initiatives**. While non-legally binding, they have nevertheless the effect of prompting private actors into the development of measures, e.g., in the form of **codes of conduct**. Their implementation may also entail the removal of certain types of online content. For example, it has been reported how in the UK search engines have agreed to a 'Voluntary Code of Practice' that requires them to take

[32] Under the German legislation, persistent failure by large social networks (e.g., Facebook) to remove within 24 hours illegal online content can result in fines of up to 50 million EUR.

[33] Mchangama, J. and Fiss, J. (2019), 'The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship', *Justitia*.

[34] See David Kaye, former Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression', 1 June 2017, https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf . In turn, at least three countries – Russia, Singapore, and the Philippines – have directly cited the German law as a positive example as they contemplate or propose legislation to remove illegal content online. See Human Rights Watch (2018), 'Germany: Flawed Social Media Law – NetzDG is Wrong Response to Online Abuse', https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

[35] LOI n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet.

[36] Conseil constitutionnel, Décision n° 2020-801 DC du 18 juin 2020, Loi visant à lutter contre les contenus haineux sur internet. In its decision, the French Constitutional Council quashed provisions of the law that introduced: a one hour time limit t for the intermediary's deadline to remove illegal terrorist content and child sexual abuse after the receipt of a notification by an administrative authority; the so-called 'best-efforts obligations' linked to the removal measures such as transparency obligations (in terms of access to redress mechanisms and content moderation practices, including the number of removed content, the rate of wrong takedowns,…); as well as the oversight mandate the law entrusted to the French High Audio-visual Council to monitor the implementation of those best-efforts obligations. See La Quadrature Du Net (2020), 'Loi haine : le Conseil Constitutionnel Refuse la Censure sans juge', https://www.laquadrature.net/2020/06/18/loi-haine-le-conseil-constitutionnel-refuse-la-censure-sans-juge/.

[37] Breyer, P. (2020), 'French Law on Illegal Content Online Ruled Unconstitutional: Lessons for the EU to Learn', https://www.patrick-breyer.de/en/french-law-on-illegal-content-online-ruled-unconstitutional-lessons-for-the-eu-to-learn/.

'additional steps' to remove links to allegedly unlawful content[38]. This type of measures have been referred to as forms of 'shadow regulation'[39].

Content moderation practices enforced on the basis of— direct or indirect— government or public regulators' demands are increasingly coupled with those enforced as a result of private actors' self-regulation. Self-regulation has been defined as 'the regulation of a company or a sector of itself in order to achieve an industry *or* public policy objective'[40], without however acting upon direct or indirect obligations or pressure from public or state regulators. In a context where user-generated content is not subject to the editorial controls associated with traditional media, **self-regulation is increasingly viewed as a solution to enhance online accountability and reduce the risk of harm that certain types of information or message can cause to online users**. There are, however, examples where content moderation through self-regulation is functional to the pursuit of goals that are of purely commercial nature. These range from addressing spam to the de-prioritisation of user-generated content and prioritisation of other type of contents (e.g., traditional media) that are considered 'more predictably advertiser-friendly'[41].

I.1. CONTENT MODERATION THROUGH ONLINE PLATFORMS' INTERNAL RULES AND POLICIES

Whether the result of co-regulation or the outcome of self-regulatory efforts, moderation of online content increasingly relies upon and stems **from the implementation of online platforms' in-house policies and related enforcement practices**. Under co-regulatory and self-regulatory schemes, online content controls and restrictions are largely imposed through the internal rules of the online platforms. The naming of these internal rules varies from company to company, and so does their content, as well as the approaches and techniques adopted by service providers in regulating user's behaviours.

In the first place, online platforms' (co or self)-regulation of online content moderation entails the identification of **what is to be considered as prohibited content or disallowed speech**. As already mentioned, such identification depends on **what specific laws oblige social networks and online platforms to consider 'manifestly unlawful' or 'harmful'**. The United Nations Human Rights Office has underlined that **poor and vague definitions of what constitutes "unlawful or**

[38] McSherry, C., York, J.-C. and Cohn, C. (2018), 'Private Censorship Is Not the Best Way to Fight Hate or Defend Democracy: Here Are Some Better Ideas', https://www.eff.org/el/deeplinks/2018/01/private-censorship-not-best-way-fight-hate-or-defend-democracy-here-are-some.

[39] Association for Progressive Communication (2018), 'Content Regulation in the Digital Age: Submission to the United Nations Special Rapporteur on the Right to Freedom of Opinion and Expression', p. 9.

[40] Council of Europe (2021), op. cit., p. 39. Emphasis added.

[41] Caplan, R. and Gillespie T. (2020), 'Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy', *Social Media + Society*, Vol. 6, No 2, p. 1.

harmful content" are among the **main key cross-cutting challenges** characterizing online content laws.

Under EU law, for instance, there are currently five types of content that are deemed unlawful, namely: child sexual abuse material[42], racist and xenophobic hate speech[43], commercial scams and frauds[44], as well as content infringing intellectual property rights[45], and terrorist content[46]. International human rights bodies and civil society organisations have underlined the **inherent legal imprecision and unclarity characterising some of these EU definitions which tend to be too broad or over-inclusive**, endangering legal certainty and disproportionately affecting protected rights and freedoms[47].

Qualifications of what constitutes illegal or 'harmful' content frequently depends on **content-specific restrictions that governments often introduce in reaction to certain events** (e.g., web streaming of terrorist attacks) or **trends** (e.g., increase in racist speech), and in response **to perceived public pressure to act**. International human rights bodies have recently highlighted how certain governments around the world have viewed legislation targeting online content as an **effective means to limit speech they dislike or to silence civil society or other critics**[48].

In a context where different norm-makers and regulators (at the national, regional, and international venues) adopt or promote their own definitions of what is illegal or harmful content, **online platforms are increasingly obliged to create** ***country-specific or jurisdiction-***

[42] Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography and replacing Council Framework Decision 2004/68/JHA, OJ [2011] L335/1; see also European Commission Proposal for a Regulation laying down rules to prevent and combat child sexual abuse, COM(2022) 209 final, Brussels, 11 May 2022.

[43] 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, OJ [2008] L328/55.

[44] Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as regards the better enforcement and modernisation of Union consumer protection rules, OJ [2019] L328/7.

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

[46] Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA,OJ [2017] L88/6; and Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online of 29 April 2021, and the definition of 'terrorist content' laid down in Art. 2.7.

[47] United Nations Special Rapporteur's Comment on 'A Counter-Terrorism Agenda for the EU: Anticipate, Prevent, Protect, Respond', COM(2020) 795, 21 October 2021, OL OTH (229.2021) (ohchr.org). See also UN Special Rapporteur on Free of Expression and Opinion, 'Comment on draft EU Regulation on preventing the dissemination of Terrorism Content Online', 3 November 2020, https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=25661; as well as, for instance, the response by European Digital Rights (EDRI) and 114 civil society organisations of the above-mentioned 2022 Commission Proposal for a Regulation laying down rules to prevent and combat child sexual abuse. EDRI (2022), European Commission: uphold privacy, security and free expression by withdrawing new law, 8 June 2022, https://edri.org/wp-content/uploads/2022/06/European-Commission-must-uphold-privacy-security-and-free-expression-by-withdrawing-new-law.pdf.

[48] OHCHR (2021), 'Moderating online content: fighting harm or silencing dissent?, 23 July 2021.

specific online content moderation policies. At the same time, the scope and possible targets of content moderation can also constitute the result of the interplay or overlaps of multiple regulatory and policy frameworks: e.g., those in force in the online platform's country of establishment, those provided in the different countries where the company operate, and those created by international cooperation, through harmonisation at the supranational level, for instance between EU Member States.

In such a complex normative and policy context, there exists the risk that online platforms providers become vehicles or proxies of both over-implementation and over-enforcement that impact fundamental rights and freedoms. There is no legal certainty as to the exact scope and content of their obligations due to the multiplicity of applicable rules, which may also hinder their effectiveness.

In addition to the multiplication of laws and policies regulating online content, private companies have also developed and instituted their own rules to set out what is to be considered 'prohibited content' on their platforms. Online platforms have increasingly engaged in the development of content moderation rules and strategies that apply regardless of *where* certain content is, for instance, produced or uploaded, or of *where* the user producing or disseminating such content sits. While the modalities for content moderation deployed by online platforms vary significantly (in terms of scope of application, organisational architecture, enforcement mechanisms), the main goal of these internal rules is to provide the basis for content moderation practices that are not country and/or jurisdiction-specific, and which are thus 'scalable' to the cross-jurisdictional nature of the services provided by online platforms[49].

It has been observed that in a situation where online platforms actively contribute to shape the scope and target of content moderation through their internal rules, content restrictions based purely on internal rules by different service providers become '*de facto* global standards'[50]. These standards are enshrined in and enforced through different sets of content moderation policies developed by online platforms, which include different sets of documents ranging from Terms of Service (ToS), community guidelines, as well as internal documents for human moderators.

Online content moderation rules set by online platforms are first provided under the Companies' Terms of Use, Terms of Services, or Terms and Conditions (ToS), which constitute the legal document a person must agree to abide by when registering an account with an online platform. ToS provide the basis on which online platforms perform content moderation, and in particular: first, to assess the (il)-legality or (in)-compatibility with terms of service of third-party content; and second, to decide whether certain content posted, or attempted to be posted, should be removed or left online, rendered less accessible, tagged as being potentially

[49] De Streel, A. (2020), 'Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform', Study by the IMCO Committee, European Parliament, PE652.718, p. 51.

[50] Council of Europe (2021), op. cit., p. 17.

inappropriate or incorrect, or demonetised. In other words, ToS can potentially implement a wide range of content restrictions via enforcement of private moderation practices.

From a legal perspective, **ToS constitute standardised contracts**. As such, they are defined unilaterally by online platforms and offered on equal terms to any user. ToS are part of the legal category of 'adhesion agreements', under which users do not have the choice to negotiate but can only accept or reject the contract terms. In fact, it has been observed that **these agreements establish a kind of 'take it or leave it' relationship which replaces 'the traditional concept of bargained clauses among contracting parties'**[51].

Furthermore, online platforms often draft community guidelines. These are non-legally binding documents developed to set out which content and behaviours are deemed 'unacceptable', 'disruptive' or 'inappropriate', as well as to direct the behaviour of all platform users. Community guidelines are typically designed to be 'living documents' and are, consequently, constantly evolving. Community guidelines can be conceived in-house by social media companies, but they may also arise from different types of consultations of the online community or through the work of stakeholders' engagement teams which engage with scholars and/or civil society organisations representing users that might be impacted by the company's online content moderation policies[52].

Research has concluded that throughout the last decade there has been **a clear trend across the largest platforms towards greater complexity in their content moderation policies**[53]. Nowadays, content moderation policies (as defined in ToS and/or community guidelines) contain specific provisions on various content categories spanning multiple web pages, blog posts, etc. Within these documents, the rights and responsibilities between the parties are allocated, users are instructed ('educated') to communicate in a civil and non-harmful manner, informed about potential monitoring and moderating measures by the online platform, and about the consequences they face when violating the rules.

Yet, a wealth of research and the discussions during the various Task Force meetings demonstrate that **content moderation policies defined in ToS and/or community guidelines are often phrased vaguely and too broadly**[54]. For instance, research has found that major online platforms' definitions of what constitutes 'hate speech' is only partially in line with the one

[51] Venturini J., Louzada L., Maciel M.F., Zingales N., Stylianou K. and Belli L. (2016), *Terms of Service and Human Rights: An Analysis of Online Platform Contracts*, Editora Revan.

[52] See for instance, Twitter Safety (2021), 'Our Continued Collaboration with Trusted Partners', Twitter Blog, 17 December 2021, https://blog.twitter.com/en_us/topics/company/2021/our-continued-collaboration-with-trusted-partners; Meta (2022), 'Bringing Local Context to Our Global Standards, Meta Transparency Center, 28 January 2022 (updated), https://transparency.fb.com/policies/improving/bringing-local-context.

[53] For a quantitative account in this regard, see Díaz, A. and Hecht-Felella, L. (2021), *Double Standards in Social Media Content Moderation*, Brennan Center for Justice, p. 5.

[54] Venturini J., Louzada L., Maciel M.F., Zingales N. and Stylianou K., Belli L. (2016), op. cit., p. 24 and 54. See Einwiller, S.-A. and Kim, S. (2020), 'How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation', Policy & Internet, Vol 12, no. 2, p. 184-206.

provided by relevant international human rights Instruments, most notably in Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR)[55]. In particular, it appears that **community guidelines of several companies extend the notion of 'hate speech' to forms of expression (e.g., 'harmful stereotypes') that fall outside the spectrum of speech that is prohibited under international human rights law** instruments, or that do not strictly limit the grounds upon which restrictions might be justified to the characteristics protected under such instruments[56].

In this regard, Fionnuala Ní Aoláin, the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, expressed her concerns regarding the Facebook community guidelines on dangerous individuals and organisations[57]. Those guidelines include commitment to remove content produced by a wide range of groups, including not only terrorism or hate speech perpetrators, but also content produced by violent non-state actors, which are organisations that advocate violence against state, but not civilian, militaries, and even entities which do not incite or advocate violence, but are viewed as demonstrating an intent to do so. The Special Rapporteur stressed that Facebook's designation of violent content is vague, imprecise, and leads to criminalisation of speech[58]. While some may find some speech offensive, distasteful or unacceptable, it is in fact protected by the right to free expression under Article 19 of the ICCPR.

Furthermore, both ToS and community guidelines are **regularly subject to changes in terms of content moderation policies** that are often 'disjointed, unclear, or limited' to addressing the narrow issue arising from a specific (often high level) controversy[59]. New rules are usually announced in company statements 'spread across multiple locations', ranging from company blogs to corporate social media accounts, to third-party websites[60]. This can make it **difficult for users to properly understand and adhere to company policies.** The Task Force discussions revealed that the ever-changing nature or volatility of ToS can also **effectively hinder regulators, oversight bodies and civil society actors' capacity to hold the platforms accountable for their enforcement decisions.**

---

[55] Article 20(2) of the ICCPR states that 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law'. For the relationship between Article 20(2) and Article 19 in the fight against online hate speech see Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'Online Hate Speech', 9 October 2019, A/74/486, para. 12. In the report it is recalled that the Rabat Plan of Action (RPA), which assists in the interpretation and application of Article 20(2) of ICCPR, stresses that this provision needs to be read in light of Article 19.

[56] Mchangama, J., Alkiviadou, N. and Mendiratta, R. (2021), 'A Framework of First Reference: Decoding a human rights approach to content moderation in the era of "platformization"', Justitia Report, November 2021.

[57] Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism (2021), 'Input of the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism to the Facebook Oversight Board Concerning its "Community Guidelines" and "Community Standard on Dangerous Individuals and Organizations"'.

[58] Ibid, p. 2.

[59] Díaz, A. and Hecht-Felella, L. (2021), op. cit., p. 5.

[60] Ibid.

Platforms retain considerable discretion regarding content they can choose to act on. It has been noted how the unclear or open-ended formulation of internal content moderation rules and policies allows internet intermediaries 'maximum flexibility' to act if they are put under pressure by, for example, a state[61]. In practice, this flexibility can even enable re-interpretation of certain provisions, with cases reported where similar content was treated differently on a case-by-case basis[62].

In addition to public community guidelines, online platforms maintain internal documents which guide the human moderators employed by the companies[63]. **Internal documents** for moderators, which have been described as 'a much more in-depth version of community standards', are mainly concerned with enabling and regulating the active monitoring and moderation of user-generated content. While the details of such documents are usually not shared (e.g., to avoid moderation practices to be circumvented) their **lack of publicity contributes to the opacity and lack of accountability and transparency** of the internal processes of content regulation by private platforms.

I.1.1. Online content moderation through automated decision-making

Online platforms combine human moderation of online content with automated decision-making processes that perform a wide range of online content surveillance and moderation tasks through **the deployment of various Artificial Intelligence (AI) systems**.

Currently, there are different ways in which AI is used for content moderation purposes. This includes spam detection, hash-matching technology— which entails the use of digital fingerprints to identify, for instance, terrorist or child exploitation content—, keyword filters, and various forms of detection algorithms, which assess the nature of the content for prohibited words or imagery.

Content moderation can be fully automated, for example to directly detect and prevent the posting of content (e.g., images) violating laws (e.g., copyright or intellectual property laws). In other cases, the deployment of AI tools can be combined with human moderation (e.g., in the domains of hate speech or terrorist content). This **hybrid approach typically involves the automatic identification or pre-filtering of content that is then checked by a team of human moderators.** Hybrid forms of content moderation deployed by online platforms have been

[61] Council of Europe (2021), op. cit., p. 17.

[62] Tobin, A., Varner, M. and Angwin, J. (2017), 'Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up', *Propublica*, 28 December 2017.

[63] These can include different groups of operators, including not only the platform's employees who set the rules and oversee their enforcement, adjudicate hard cases, and influence the 'philosophical approach' that the platforms take to govern content, but also freelancers who work on contracts with the company and guard against infractions on the front line. See Roberts, S.-T. (2016), 'Commercial Content Moderation: Digital Laborers' Dirty Work', Media Studies Publications. At the same time, content moderation can also be performed by a user (e.g., when they flag a specific content), or any other 'trusted notifier;. Schemer, S-F (2019), 'Trusted Notifiers and the Privatization of Online Enforcement', Computer Law & Security Review, Vol. 35, No 6, p. 1.

described as 'a complex socio-technical system, including a multi-stage combination of human and machines that interact in complex ways'[64].

As the presentations and discussions of the Task Force revealed, the increasing deployment of algorithmic content moderation is often justified by the argument that relying exclusively on human review is unfeasible, due to the scale at which the companies operate and the volumes of data to be controlled. Furthermore, heightened recourse to automated tools has been driven by increasing governmental and supranational regulators demands upon platforms to swiftly and rapidly remove illegal and/harmful content pursuant to standing regulatory mandates. At EU level, the European Commission has called upon Internet platforms to use automatic filters to detect and remove terrorist content, with human review in some cases suggested as a 'necessary counterweight to the inevitable errors caused by the automated systems'[65]. In the UK, governmental authorities have reportedly developed a tool to automatically detect and remove terrorist content at the point of upload[66].

Prior research has underscored the benefits of deploying AI systems to support human moderators in screening and analysing huge amounts of data and information, while relieving them from the psychological burden of seeing/reading harmful content[67]. Yet, to date **the quality and reliability of content moderation performed by AI tools remains limited in many respects**. Research has found that the AI 'machine struggles with short and long messages'[68]. Another challenge relates to the limitations of AI tools in assessing context of language and speech, and in taking into account 'widespread variation of language cues, meaning and linguistic and cultural particularities'[69].

Because of AI's dependence on training data (e.g., moderators' past removal decisions), automated decision-making tools happens to merely mirror choices already made by human operators. These tools have **limited capacity to understand the specific context, and sometimes the very nature, of certain types of complex forms of speech**. A clear example of the latter is content labelled as 'hate speech', the very nature of which remains contested and extremely

[64] Sartor, G. and Loreggia, A. (2020), 'The impact of algorithms for online content filtering or moderation "Upload filters"', Study commissioned by the Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies of the European Parliament, PE 657.101.

[65] European Commission, Recommendation of 1 March 2018 on measures to effectively tackle illegal content online, C(2018) 1177.

[66] Home Office, 'New technology revealed to help fight terrorist content online', press release, 13 February 2018 https://www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online.

[67] Riedl, M. J., Masullo, G. M. and Whipple, K. N. (2020), 'The Downsides of Digital Labor: Exploring the Toll Incivility Takes on Online Comment Moderators', *Computers in Human Behavior*, Vol. 107.

[68] Ruckenstein M. and Turunen, L.L.M. (2020), 'Re-humanizing the Platform: Content Moderators and the Logic of Care', *New Media & Society*, Vol. 22, No 6, p. 1038. The authors note: 'The short messages do not have sufficient relations between words for calculations, while negative content can disappear from long messages, because the machine counts how negative a message is overall; a single harmful or inappropriate element — quickly picked up by a human reviewer — is not enough to alert the machine'.

[69] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 18-14238* (E) 301018, p. 8.

context dependent. Since algorithms still lack the capacity to perform the complete contextual analysis which is necessary to assess whether certain types of content should be restricted, content moderation systems that rely heavily or exclusively on AI tools are more likely to block or restrict content by default. This, in turn, carries the risk of removal of online content that is not unlawful, or the suspension of accounts that are not problematic[70].

Furthermore, because AI applications are often grounded in datasets that incorporate discriminatory assumptions[71], the risk exists that the algorithms used to operate **AI-driven content moderation reproduce and amplify biases** of those involved in their development and/or in the development of the algorithms' training data or of prior decision making that feeds into the algorithms. Researchers have found **racial biases in some AI hate speech detectors**, with cases being reported of content removed in accordance with biased or discriminatory concepts, and **structurally vulnerable groups**—such as communities of colour, LGBTQ+ communities, religious and ethnic minorities, etc.— having been found **most likely to be disadvantaged** by AI content moderation systems[72].

Such findings call for a serious reconsideration of **whether, how, and the extent to which automated decision making tools should be used to perform moderation** of certain types of online content. Several legality and accountability challenges arise from the increasing reliance by online platforms on AI to make highly complex and sensitive decisions (e.g., about the lawfulness of certain content, or its compliance with terms of service) which have **far reaching impact for the fundamental rights of the concerned individuals, but also on society at large**[73].

At EU level, Article 22(1) of the General Data Protection Regulation (GDPR) provides data subjects with **the right 'not to be subject to a decision based solely on automated processing, including profiling**, which produces legal effects concerning him or her or similarly significantly affects him or her'. However, the interpretations of what 'significant effects similar to legal effects' the potential occurrence of which should prevent individuals from being subject to algorithmic decisions vary significantly from Member State to Member State[74]. The specific interpretation given to this provision in the EU27 may therefore impact the margin of

[70] Ibid.

[71] Ibid.

[72] Díaz, A. and Hecht-Felella, L. (2021), op. cit.

[73] See UN Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, Report on Racial discrimination and emerging digital technologies: A human rights analysis, 18 June 2020, A/HRC/44/57. Nahmias, Y., Perel, M. (2021), 'The Oversight Of Content Moderation By Ai: Impact Assessments And Their Limitations', Harvard Journal on Legislation, Vol. 58, p. 150. The authors note: 'The application of AI-based content-moderation systems by prominent online platforms is riddled with externalities. It directly affects people's ability to engage in certain forms of expression, communication, and sharing of thoughts and critical information. Consequently, it shapes our online public sphere and ultimately governs the free flow of information'.

[74] Maugeri, G. (2019), 'Automated Decision-making in the EU Member States: The Right to Explanation and Other 'Suitable Safeguards' in the National Legislations', *Computer Law and Security Review*, Vol. 35, No 5.

manoeuvre social platforms have to resort to AI systems in order to perform online content moderation.

The GDPR provides for explicit exceptions to the prohibition on fully automated individual decision-making. These include, *inter alia*, cases where this type of processes/decisions are authorised by EU law, or by the Member State law to which the data controller is subject. In such cases, **the law must lay down adequate measures to safeguard the data subject's rights, freedoms, and legitimate interests**. In this regard, Article 22(3) of the GDPR establishes that individuals subject to (authorised) forms of fully automated decision-making (including in the online content moderation domain) have the right to express their point of view, the right to obtain human intervention on the part of the controller, as well as the right to contest the decision.

This set of safeguards is usually referred to as the 'right to explanation'. Such right is the object of a persistent debate in legal doctrine. While some scholars went as far as doubting its very existence under the GDPR[75], others have interpreted it as requiring the assurance of effective oversight mechanisms that, beside the availability of remedies to be invoked by data subjects on an individual basis, should also cover the design, prototyping, testing, and practical deployment of data processing systems[76].

Despite the (limited) guidance provided by certain national Data Protection Authorities (DPAs) on the ways in which data controllers should implement the guarantees pertaining to the right to explanation in the context of fully automated decision-making[77], **the exact content of the information which controllers need to give data subjects about the algorithms' underlying logic, significance and envisaged consequences for the individual has not yet been completely clarified**[78].

At present, online platforms do not systematically produce reports providing details regarding how much content is restricted, and for which specific reasons through automated decision-making systems. They also don't publish information regarding the specific reasons why certain content was removed (or not) through the use of AI systems. **AI-driven moderation of content remains particularly opaque.**

[75] Edwards, L. and Veale, M. (2017), 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably not the Remedy You Are Looking for', *Duke Law & Technology Review*, Vol. 18, p. 16.

[76] Nahmias, Y. and Perel (Filmar), M. (2021), 'The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations', *Harvard Journal on Legislation*, Vol. 58, No 1.

[77] Maugeri, G. (2019), op. cit.

[78] Barros Vale, S. and Zanfir-Fortuna, G. (2020), 'Future of Privacy Report: Automated Decision-Making Under the GDPR – A Comprehensive Case-Law Analysis', May 2022. The authors note how a recent preliminary ruling request sent by an Austrian court in February 2022 to the Court of Justice of the European Union (CJEU) may soon clarify these concepts, as well as others related to the information which controllers need to give data subjects about the underlying logic, significance, and envisaged consequences for the individual of automated decision-making.

It is well known that automation can impede the transparency of a decision-making process and hinder the possibility to provide an explanation of how a certain outcome or decision has been reached. The Task Force discussions and presentations highlighted **the need for more precise guidance from regulators at national and supranational (EU) levels,** including about the transparency requirements that online platform providers should meet in order to deploy automated content moderation. In this regard, the GDPR requires entities implementing AI-based or other types of automated decision-making systems that are likely to result in high risk to an individual rights and freedoms **to carry out a data protection impact assessment prior to their adoption**[79]. However, the type of impact assessments required by the GDPR have been subject to criticisms insofar as they only provide limited transparency, insufficiently secure due process, and only allow limited room for public review[80].

The risk therefore is that that the public space monitored by social platforms content moderation algorithms may be different for everyone. In this regard, it has been observed that '[a]scertaining whether an AI system impacts human rights, democracy and the rule of law can be rendered difficult or impossible when there is no transparency about whether a product or service uses an AI system, and if so, based on which criteria it operates. Further, without such information, a decision informed or taken by an AI system cannot be effectively contested, nor can the system be improved or fixed when causing harm'[81]. Consequently, **there is often no way of meaningfully knowing whether some discriminatory results occur as a result of AI operations in the online content moderation domain**.

I.1.2. Multilateral cooperation fora

Content moderation by online platforms is also developing through multilateral fora aimed at increasing private sector coordination in the fight against the dissemination of illegal or harmful content online. Recent years have been marked by the proliferation of arrangements between platforms to work together to remove certain categories of content or actors from their services. These arrangements are increasingly proposed as a response to a wide number of challenges in the online environments.

---

[79] See GDPR, Article 35. Article 35(3)(a) establishes that this is particularly the cases when the case when the data controller uses systematic and extensive evaluation of "personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person. The GDPR further states that in order to help data controllers ascertain whether processing is likely to present high risks, supervisory authorities will maintain a list of processing operations which are subject to the requirement for DPIA, or for which no impact assessment is required. See, Article 35 (4)-(5) Ibid.

[80] Nahmias, Y. and Perel (Filmar), M. (2021), op. cit., p. 8. The authors state that the kind of impact assessment mandated under the GDPR is tailored to mitigate concerns about the ways general AI-based decision-making systems affect individuals. They are not, however, designed to address the consequences of poorly performed AI driven content-moderation on the online public sphere. Ibid, p. 7.

[81] Council of Europe Ad Hoc Committee on Artificial Intelligence (2020), 'Feasibility study', CAHAI(2020)23, 17 December 2020, p. 33.

Child Sexual Abuse Material (CSAM) has been the first domain where cooperation between large online platforms was developed, in particular through the creation of databases to which certain online content (and more specifically, images that are 'hashed' through various forms of digital fingerprinting technologies) are added. Companies can then use these databases to prevent copies of included images automatically and pre-emptively from being uploaded to their services[82]. Civil society actors have long been vocal about **the opacity and risks associated with this model of centralised private censorship**[83].

Similar types of private actors' cooperation fora have proliferated to address various other online safety issues, as demanded or encouraged by regulators. In relation to 'terrorist content', Facebook, Google, Microsoft, YouTube, and Twitter cooperate through the Global Internet Forum to Counter Terrorism (GIFCT). The aim of the forum is to curtail the spread of terrorism and violent extremism through technical solutions, research, knowledge sharing, and building of the Shared Industry Hash Database. These providers share best practices for developing their automated systems and operate a 'hash database' of terrorist content, where digital fingerprints of illicit content (images, video, audio, and text) are shared[84]. Despite early concerns raised by certain internet industry representatives about the possibility to safely deploy hashing technology to pre-emptively check uploads to platform services and prevent posting of content in the sensitive and controversial domain of terrorist speech[85], this database is currently used to track and share content that at least one participating company has identified as 'terrorist propaganda' (regardless of whether their criteria align).

Though voluntary in nature, the GIFCT is part of **a commitment to increase industry collaboration that has also been compelled by regulators**. In particular, it constituted a response to the European Commission's request for online platforms to adhere to the 2016 Code of Conduct to combat illegal online hate speech[86]. The Code of Conduct requires private actors to review the majority of hateful online content within 24 hours of being notified, and to remove it in the name of combating hate speech and terrorist propaganda across the EU. Despite not legally binging instrument, **the Code of Conduct has *de facto* put more responsibility upon platforms to police content**, without the accountability and oversight of democratic institutions, or the inclusion of specific substantive and procedural safeguards to

[82] Doudek, E. (2020), 'The Rise of Content Cartels', Knight First Amendment.

[83] EDRI (2012), The Rise of the European Upload Filter, 20 June 2012.

[84] Within hours of the Christchurch attack, Facebook had uploaded hashes of about 800 different versions of the shooter's video. Technical and computational feats enable every single video and image uploaded by ordinary Facebook users (as well as YouTube and Twitter users) to be 'hashed' and checked against the database. If it matched, it would be blocked. See, Sonderby, C., (2019), 'Update on New Zealand', Facebook Newsroom, https://perma.cc/ZA85-2Y3X.

[85] Doudek, E., (2020), op. cit.

[86] European Commission (2019), Assessment of the Code of Conduct on Hate Speech on line State of Play, Council of the European Union Document 12522/19, 27 September 2019.

ensure that lawful content (for example, journalism covering the topic of extremism) is not arbitrarily taken down[87].

At global level, cooperation between platforms and government is expanding to further areas (e.g., foreign influences on election campaigns) where there is convergence of private and public interests to tackle the circulation of content across the entire internet ecosystem[88]. In the UK, calls for further cooperation among companies have been made in the White Paper on online harm, which stressed that a 'greater level of cooperation between platforms by sharing observations and best practices to prevent harms spreading from one provider to another will be essential'[89].

In parallel, concerns have been raised during Task Force meetings concerning **the structural lack of external oversight of industry cooperation in online content policing**. To date, the database operated by the GIFCT remains secretive. There are no independent mechanisms to audit or challenge inclusion of content in such a system. This type of cooperation has been criticised as compounding the lack of legitimacy, transparency, due process, and accountability already affecting content moderation by online platforms[90].

Calls from representatives of civil society organisations have been made to address critical shortcomings affecting content moderation, including most notably **the lack of transparency related to the scope, scale, and impact of such practices, and the absence of robust, transparent, and reliable mechanisms for due process available to users**. Illustratively, the 'Santa Clara Principles on Transparency and Accountability in Content Moderation' is a civil society-driven initiative developed to promote industry convergence around a set of minimum standards for fairer, unbiased, proportional online platforms' content moderation, and to support companies to comply with their responsibilities to respect human rights and enhance their accountability, and to assist human rights advocates in their work[91].

---

[87] Huszti-Orban, C., (2017), op. cit.

[88] Brandt, J. and Hanlon, B. (2019), 'Online Information Operations Cross Platforms. Tech Companies' Responses Should Too', *Lawfare*, 26 April 2019.

[89] Secretary of State for Digitalisation, Culture, Media and Sport and Secretary of State for the Home Department (2019), Online Harms White Paper, 45.

[90] Llansó, E. (2019), 'Platforms Want Centralized Censorship. That Should Scare You', *Wired*, 18 April 2019. In this regard, see also Doudek, E. (2020), op. cit.

[91] These may be found here: https://santaclaraprinciples.org/. Released in 2018 and opened for revision in 2020, the principles have been adhered to by major online platforms — including Apple, Facebook (Meta), Google, Reddit, Twitter, and GitHub. They are also addressed at smaller, newer, and less resourced companies to inform future compliance. The principles include specific guidance about what information is needed to ensure meaningful transparency and accountability in private-led content moderation. While the principles are reportedly not designed to provide a template for regulation, they include recommendations for governments and other state actors to respect their international human rights law obligations and not to exploit or manipulate private online moderation systems for censorship. They also recommend governments to remove barriers for company transparency about how they restrict content online. See Llansó, E. (2021), 'Santa Clara Principles 2.0: Civil Society Recommendations for How Companies, States Should Protect Free Expression Rights Online', Center for Democracy and Technology Insight–s - Free Expression, 8 December 2021.

I.2. CONTENT MODERATION AND INTERNAL OVERSIGHT BY ONLINE PLATFORMS

By contrast to content removal measures executed in response to a mandatory order issued by a public (i.e., judicial or administrative) authority, **online content-related decisions taken by online platforms pursuant to the terms of service constitute the outcome of purely internal assessments**. Content moderation actions — e.g., about whether to host or continue hosting a specific piece of content, or to take it down permanently or temporarily, either on the platform as a whole or in relation to certain users in a specific geographical area — to enforce platform's ToS are adopted internally and unilaterally. **They are carried out without independent judicial or administrative oversight.**

The impact that these content moderation practices have for the rights of concerned individuals have led representatives of international human rights organisations[92] and civil society actors[93] **to recommend the establishment of internal (i.e., non-administrative and non-judicial) accountability mechanisms for users to report harms and raise concerns related *inter alia* to freedom of expressions and privacy**. This type of internal accountability mechanisms for the private sector is also called for by the **United Nations Guiding Principles on Business and Human Rights**[94]. In the content moderation domain, they can take the form of multi-stakeholders' advisory boards mandated to review private decisions on content moderation, and to serve communities against private infringement of rights and freedoms.

Lack of clear and accessible mechanisms for users to submit grievances and concerns related to fundamental rights violations potentially deriving from private companies' content moderation decisions has reportedly been a long-standing feature of many companies operating in internet ecosystem, including certain large online platforms. As a way of illustration, a 2019 evaluation of the mayor online platforms' 'grievance and remedy' mechanisms concluded that Google did not offer individuals affected by content moderation decisions the possibility to appeal and provide further information. The company only gave options for users to appeal certain actions that could impact freedom of expression or privacy, such as copyright takedown decisions, account restrictions, or sharing user data. It was unclear if users could submit complaints about other types of actions that a user felt infringed on their

[92] Kaye, D. (2019), 'An Open Letter to Mr. Zuckerberg', Office of the United Nations High Commissioner for Human Rights, 1 May 2019.

[93] Article 19 (2019), 'The Social Media Councils: Consultation Paper'. See Also, Access Now (2019), 'Protecting free expression in the era of online content moderation: Access Now's preliminary recommendations on content moderation and Facebook's planned oversight board', p. 2.

[94] Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework. The guiding principles were developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises. The Special Representative annexed the Guiding Principles to his final report to the Human Rights Council (A/HRC/17/31), which also includes an introduction to the Guiding Principles and an overview of the process that led to their development. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.

freedom of expression or privacy. The evaluation also noted how the company offered hardly any evidence related to follow up to these complaints[95].

I.2.1. Online platforms' internal accountability through self-regulation

Most recently, most social media platforms have equipped themselves with **internal mechanisms of complaint against decisions made about users' content**. The structure, mandate, procedural settings, and enforcement options of these mechanisms **differ significantly from company to company**[96]. An important experiment undertaken by global online platforms in this domain is provided by the **Facebook Oversight Board** (FOB)[97]. The creation of the Oversight Board was anticipated by findings of critical gaps in Facebook's internal 'grievance mechanisms'[98]. The new private body was established by Facebook in 2020, following a series of exchanges with external stakeholders[99].

The Board's competences and scope, composition, process for the selection and removal of members, operating procedures, and governance models are set out in the FOB's Charter[100], as complemented by its Bylaws[101]. According to the latter, the FOB admits submissions of users' requests for review which encompass claims for reinstatement of removed posts. The final choice about which requests the FOB will review is made by the Board's own Selection Committee, which decides based on criteria that include the difficulty and significance of the case. Since the Board started accepting cases in October 2020, more than half a million users submitted appeals to it[102]. Out of these, the Oversight Board has reportedly taken on 21 high-level cases (i.e., cases that the FOB assessed as difficult and significant), and proceeded with

[95] See findings from the 2019 Ranking Digital Rights Corporate Accountability Index. Ranking Digital Rights (2019), '2019 RDR Corporate accountability Index', Section 3–3 - Grievance and remedy.

[96] The Electronic Frontier Foundation attempted to develop a synthetics comparative overview of the options users have to complain and of how decisions are made inside different social media company. Electronic Frontier Foundation, 'Tracking global Online Censorship: Tracking the impact of content moderation on freedom of expression worldwide'.

[97] Klonick, K. (2020), 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression', Yale Law Journal, Vol 129, No. 2418, 2020. The authors notes that at the basis of such calls there is the increasing public awareness of censorial control of large private platforms, and the consequent demands for greater accountability and transparency to users about how content moderation decisions are made.

[98] Ranking Digital Rights (2019). According to the 2019 ranking, Facebook mechanisms were among the weakest of any company in the RDR Index. It noted that the company lacked a clear appeal mechanism allowing users to seek remedy in cases where they feel that Facebook has violated their privacy.

[99] Facebook (2019), 'Global Feedback and Input on the Facebook Oversight Board for Content Decisions'. Throughout the process that led to the establishment of the Oversight Board, critical recommendations were made regarding the legal standards to follow for the operation of such body, and the levels of transparency and independence required to perform its tasks. See for instance, Kaye, D. (2019), 'An Open Letter to Mr. Zuckerberg', Office of the United Nations High Commissioner for Human Rights, May 1, 2019.

[100] Facebook Oversight Board (2019), Facebook Oversight Board Charter.

[101] Facebook Oversight Board (2022), Oversight Board Bylaws.

[102] Oversight Board transparency report–s - Q4 2020, Q1 & Q2 2021.

17, which involved a variety of issues (e.g., hate speech, nudity, content by dangerous individuals/organizations, and Covid-19 disinformation).

The FOB's decisions are adopted by a panel of five members (which must include at least one representative from the region implicated), although inputs might also be provided by external experts, organisations, as well as Facebook itself. After a period of public commentary during which a brief description of the case is published, the decision adopted by the panel is submitted to the rest of the FOB for review and approval. Decisions may include policy recommendations to which the company should respond to, as well as resolutions on cases that the FOB bylaws define as binding[103]. On both content decisions and policy advice, Facebook is requested to publicly disclose the action it takes in response to the Board's decision.

With regard to the legal standards or norms guiding the FOB decisions, they have been found to embrace those provided under international human rights law instruments (in particular treaties, declarations, and reports from organisations in the universal international human rights system) when assessing the appropriateness of Facebook's content moderation decisions. At the same time, while it appears that the Board retains a large margin of discretion as to which specific instrument is used to decide on a specific case[104], the company's community guidelines remain the ultimate parameter against which the Board performs its assessments[105].

To a certain extent, the FOB can be positioned among self-regulation mechanisms on press or media content (e.g., media ethics boards, journalism associations)[106]. These structures and forms of evaluating decisions made by private companies are indeed recognized and encouraged at the international level[107]. As already mentioned above, the establishment by online platforms of clear and accessible grievance mechanisms enabling users to submit complaints in case they feel their rights have been violated by the company's actions or policies is recommended under the UN Guiding Principles on Business and Human Rights. These principles emphasize that in order to be effective, internal accountability procedures should be clear, accessible, predictable, and transparent[108].

However, doubts still exist as to the extent to which the Board is accessible, as well as with regard to its contribution in increasing transparency of content moderation practices

[103] Article 2, Section 2.3.1 of the FOB's Bylaws stresses that 'Facebook will implement board decisions to allow or remove the content properly brought to it for review within seven (7) days of the release of the board's decision on how to action the content. In addition, Facebook will undertake a review to determine if there is identical content with parallel context associated with the board's decision that remains on Facebook'.

[104] Mchangama, J., Alkiviadou, N. and Mendiratta, R. (2021), op cit.

[105] Article 1, Section 3 of FOB's Bylaws stresses that '[t]he board will review and decide on content in accordance with Facebook's content policies and values".

[106] Article 19 (2019), op. cit.

[107] For instance, that, at the regional level, several decisions by the European Court of Human Rights (ECtHR) incorporate or refer to professional and/or ethical decisions of Press Associations.

[108] Principle 31.

implemented by the platforms subject to its oversight. With regard to accessibility, the remit of the Oversight Board has been designed restrictively. To date, the Board can only take up appeals against removal of content and cannot (yet) look at decisions where disputed material is left up rather than taken down[109]. Furthermore, the scope of the Board's activities is also limited when removal is in compliance with the law of a country[110], which might, however, require mass removal of posts in compliance with repressive speech laws[111]. The general rule is that the Board cannot review cases that are not appealed[112].

With regard to transparency, requests from the FOB for information from the when reviewing cases can contribute to highlighting a lack of transparency in some content moderation decision-making criteria. This can lead to a better understanding of the moderation made by the platform. Nevertheless, it appears that **adverse decisions by the Board would not affect crucial aspects of Facebook's business model**, with the Board itself acknowledging for instance that Facebook has not been 'fully forthcoming with the Board on its "cross-check" system, which the company uses to review content decisions relating to high-profile users'[113]. The potential of the Board to foster transparency is also limited to the company's selection of information by algorithms, or issues related to storage and use of users' personal data[114]. Furthermore, Members of the Board cannot interact with government officials regarding their service on the Board and/or the cases that they are reviewing[115]. The wider framing of these provisions seems to indicate that the possibility of Board's members to transparently cooperate with public authorities (possibly also including regulators) has been limited by design.

Another source of concern concerns the level of independence the FOB will manage to build from the company from which it originates. Not only has Facebook reserved for itself the initial task of selecting the first members of the Board. The Task Force discussions revealed scepticism as to the power of the FOB to independently modify certain criteria in the Founding Charter. Experts also pointed out that due to its administrative costs and complexity of operation, this mechanism is not readily transferable to just any online platform.

---

[109] Article 3 Section 1.1 of the FOB Bylaws. Section 1.1.2 of Article 3 of the Bylaws foresees that in the future it will be possible for the Board to review decisions related to content reviewed by Facebook for potential violations of content policies and ultimately allowed to remain on the platform.

[110] Article 2 of the FoB's Charter.

[111] Bell, E. (2020), 'Facebook's Oversight Board plays it safe', *Columbia Journalism Review*, December 3, 2020.

[112] A user who has appealed a decision that affects them will receive notification of Facebook's final decision and, if the content qualifies, they will also receive a reference identification number for purposes of review by the FOB. If that person is not satisfied with the outcome of their appeal, they may choose to refer their case to the FOB within 15 days of Facebook's final decision.

[113] FOB (2021), 'Oversight Board demands more transparency from Facebook', October 2021.

[114] Article 2, Section 1.2 of the Bylaws stresses for instance that 'Not all content can be submitted to the board for its review due to technical and/or legal limitations. For example, some content is not technically or operationally feasible to be sent to the board; other content is not eligible to be submitted because of legal restrictions'.

[115] Article 5 of the FOB's Code of Conduct.

Concerns have been voiced about the Board's legitimacy as a non-state mechanism linked to the protection of freedom of expression. The Board has been criticised as an attempt to create a private Court of Justice (referred as Facebook's 'Supreme Court'[116]) to resolve fundamental rights matters that should be reserved for state authorities with democratic legitimacy, supranational courts, or international tribunals. While significant in enabling individuals to formally complain about a decision made internally by online platforms and to seek a certain type of redress against content moderation actions adopted by a company, Task Force contributions indicate that **this type of private and internal mechanisms cannot be expected to be the only form of accountability in response to the growing impact of online platforms, and fundamental rights and public debate**.

I.2.2. Online platforms' internal accountability through co-regulation

Governmental and supranational institutions also have a crucial role to play in ensuring that, in exercising online content moderation pursuant to their ToS or codes of conduct, platforms providers ensure appropriate oversight over of users' rights. Public authorities and regulators have indeed already started to support the emergence and operation of self-regulatory mechanisms, including by creating a legal underpinning for self-regulatory internal accountability mechanisms and procedures.

For instance, the German legislation (as dynamically interpreted by national courts) envisages that social networks performing online moderation pursuant to their ToS can prohibit hate speech that does not amount to a criminally punishable content pursuant to relevant national provisions[117], but only as long as deletion is not performed arbitrarily, and users are not barred from the service without recourse[118]. The law also requires platforms to provide the possibility to challenge the decision[119]. The law also expressly provides for the recognition, by the Ministry of Justice, of 'regulated self-regulatory agencies' that can decide on cases related to non-obviously illegal content. The role of these bodies, to be financed by online platforms themselves, would be to determine whether a given message is in violation of the law (rather than of the company's ToS) and should be removed. In such cases, the online platforms transfer the decision on the illegality of content to a state-recognised institution. Granting of recognition by the Ministry depends upon conditions such as the independence of the self-regulatory body, the expertise of the people making decisions, and on their capacity to reach a decision within 7 days. With regard to these specific provisions of the German law, civil society actors have expressed concerns that the guarantees provided for in the law are not sufficient to ensure neither the independence nor the effectiveness of the self-regulatory body.

---

[116] Divij, J. (2019), 'What Does Facebook's 'Supreme Court' Mean for the Future of Online Speech?', *The Wire*, 10 July 2019.

[117] Article 1(3) NetzDG.

[118] Kettemann, M.C. and Tiedeke, A.S. (2020), 'Back up: Can Users Sue Platforms to Reinstate Deleted Content?, *Internet Policy Review*, Vol. 9, No 2. Examples of this type of legislation are also coming from other EU Member States, including Italy and the Netherlands.

[119] Article 4(II)2 NetzDG.

In this regard, an important role has been played by national-level jurisprudence, which has progressively contributed to clarify the types of internal procedures and level of due diligence that online platforms must equip themselves with to enable revision of internal content moderation decisions. From Germany, in particular, a strong income of court cases relates to users' demands that their content is put back online by big platforms. Many of these cases relate to content that is labelled as 'hate speech' under platforms own rules/community standards/terms and conditions—but that does not qualify as hate speech under national legislations[120].

In two major cases litigated recently before the German Federal Supreme Court (BVerfG), the latter ruled that Facebook has a right to develop its own internal rules prohibiting certain types of speech and can enforce those rules by removing posts and closing accounts breaching those rules. Any deletion of unwelcome, yet lawful (i.e., not subject to criminal punishment) content, requires a contractual justification under the terms and conditions. These, in turn, are liable to justify the deletion if the provisions at issue are valid under the German Civil Code. At the same time, the company also has a legal obligation to take its users fundamental rights into account throughout the process.

The Court clarified that because of its size, Facebook has to comply with certain due process requirements, as the state would do, when restricting free speech. This means that Facebook would have to inform its users (at least *ex post*) of the removal of their content, and of its intention to block users' accounts. It would also have to inform users about the reasons for the action and give them the opportunity to respond in an appeal process. After this response the company should review its decision with the possibility to reinstate the removal of content. Failure to comply with these criteria will trigger the invalidity of terms and conditions under German law. The Court allows 'narrowly tailored' exceptions to these principles, which must be clearly set out in the general terms and conditions[121]. The German court took an approach to protect fundamental rights and strengthen the rule of law through procedural safeguards.

In the EU, a type of notice and take down procedure has been established after the decision in the Google Spain case[122]: if a data subject requires that certain posts be delisted (i.e., that they not be presented in response to searches that use the data subject's name), the search engine is obliged to comply with the request, unless prevailing reasons exist for allowing such search results. In this procedure, however, no voice in the delisting procedure is given to the publishers of the content being delisted[123].

[120] Kettemann, M.C. and Tiedeke, A.S. (2020), op. cit.

[121] These exceptional cases (exemptions) include deletions for purely technical reasons that do not relate to the content of the statement, e.g., removing posts of a user having irrevocably deleted their account already.

[122] Case C-131/12 *Google Spain SL and Google Inc v. Agencia Española de Protecciín de Datos (AEPD) and Mario Costeja Gonzalez*, Judgment of 13 May 2014.

[123] Sartor, G. and Loreggia, A. (2020), op. cit.

SECTION II. REGULATING AND CRIMINALISING ONLINE CONTENT: A POLICY AND NORMATIVE PRIORITY INTERNATIONALLY, IN THE EU AND THE UK

Section I of this Report has emphasised the proactive and reactive efforts of service providers to (co)-create standards in relation to online content moderation. Initiatives such as the Internet Forum, aimed at curtailing the spread of terrorism and violent extremism through technical solutions, research, knowledge sharing, and building of the Shared Industry Hash Database provide a prime example in this context[124]. This Section focuses on the regulatory standards adopted or proposed internationally (*Section II.1*), by the EU legislature (*Section II.2*), with emphasis on the DSA (*Section II.3*), and the UK, where the Online Safety Bill is currently discussed and scrutinised (*Section II.4*)[125].

II.1. INTERNATIONAL AND REGIONAL INITIATIVES

At the international level, a number of initiatives and calls have emerged in the past few years. The need to increase efforts to stop the spread of terrorist and violent extremist content in a transparent, accountable, and fundamental rights compliant manner has been recognised as a priority in the 2017 G20 Hamburg Leaders' Statement on Countering Terrorism[126], the 2019 G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism[127], as well as and the 2019 G7 Digital Ministers Chair's Summary[128].

In 2018, the UN Rapporteur on the promotion and protection of the right to freedom of opinion and expression expressed concerns regarding online content moderation globally[129]. While recognising the appeal of regulation, he drew attention to **the potential of over-removal of content by providers in order to avoid liability**[130]. The Rapporteur referred to **increasing demands for extraterritorial removal** of links, websites, and other content alleged to violate local law, which **raise serious concerns regarding interference with the right to freedom of**

[124] Marone, F. (2019), 'Digital Jihad – Online Communication and Violent Extremism', ISPI.

[125] For other countries see https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent. For developments in the United States, see https://www.orfonline.org/research/moderating-online-content-in-the-united-states/.

[126] G20 (2019), G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism Conducive to Terrorism (VECT), https://dig.watch/instruments/g20-osaka-leaders-statement-preventing-exploitation-internetterrorism-and-violent.

[127] G20 (2017), The Hamburg G20 Leaders' Statement on Countering Terrorism, https://www.g20germany.de/Content/DE/_Anlagen/G7_G20/2017-g20-statement-antiterroren___blob=publicationFile&v=2.pdf.

[128] G7 (2019), G7 Digital Ministers Chair's Summary, https://www.economie.gouv.fr/files/files/2019/G7/G7Num/Chairs_summary_version_finale_E NG.pdf.

[129] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (2018), 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression'.

[130] Ibid, p. 7.

expression[131]. Among his recommendations, he called for 'smart regulation', focused on ensuring company transparency and remediation to enable the public to make choices about how and whether to engage in online forums. He further called for an approach whereby states should only seek to restrict content **pursuant to an order by an independent and impartial judicial authority**, and where the proactive monitoring or filtering of content would be refrained from[132].

In October 2021, the UN Special Rapporteur on the Protection and Promotion of Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression, and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information issued a Joint Declaration on politicians and public officials and freedom of expression[133].

Among its recommendations, **the Joint Declaration called on social media companies to 'ensure that their content moderation rules, systems and practices reflect international human rights standards** including the importance of open and inclusive debate about matters of public interest, and elaborate clearly when, how and what measures may be taken against content posted by politicians and public officials'[134]. Other recommendations include the promotion of **the 'maximum possible transparency' regarding content moderation rules, systems and practices** — especially where these affect public interest content or content posted by politicians and public officials —, respect for basic due process principles, including by providing independent dispute resolution options, ideally overseen by independent multi-stakeholder bodies and ensuring that content moderation systems and practices take into account local languages, traditions and culture[135].

The Organisation for Economic Cooperation and Development (OECD) has also actively engaged with questions regarding online content moderation. It has published a toolkit that presents a novel typology of the different types of untruths that circulate on the Internet and calls *inter alia* for the development and implementation of online platform content moderation policies in a multi-stakeholder process and with independent oversight[136]. With regard to terrorism and violent extremism in particular, the OECD has also published two reports

[131] Ibid, p. 8.

[132] Ibid, pp. 19-20.

[133] 2021 Joint Declaration on politicians and public officials and freedom of expression https://www.ohchr.org/sites/default/files/2022-04/Joint-Declaration-2021-Politicians_EN.pdf.

[134] Ibid,

[135] Ibid.

[136] OECD (2019), 'Disentangling untruths online: Creators, spreaders and how to stop them'.

summarising and critically evaluating the world's top 50 online content-sharing services' approaches to terrorist and violent extremist content online[137].

It was found that **whereas transparency reports are increasingly published by providers with information about their practices, they remain uncommon and uncoordinated**. As a result, **the effectiveness of providers' measures remains difficult to observe, whilst there is a growing risk of regulatory fragmentation**. More recently, on 3 May 2022, the OECD launched its new Voluntary Transparency Reporting Framework, a web portal for submitting and accessing standardised transparency reports from online content-sharing services about their policies and actions on terrorist and violent extremist content online[138].

At regional level, in March 2018, the Council of Europe (CoE) adopted a Recommendation addressed to its 47 Member States aimed at clarifying the roles and responsibilities of internet intermediaries, such as search engines and social media[139]. In its Recommendation, the Committee of Ministers, which is the executive body of the organisation, has called on states to provide **a human rights and rule of law-based framework** that lays out the main obligations of the states with respect to the protection and promotion of human rights in the digital environment, and the respective responsibilities of intermediaries. The Recommendation has called on states to create a safe and enabling online environment where internet intermediaries, users, and all affected parties know their rights and duties, to encourage the development of appropriate self- and co-regulatory frameworks, and to ensure the availability of redress mechanisms for all claims of violations of human rights in the digital environment. Particular emphasis has been placed on **the freedom of expression of expression and the rights to privacy and protection of personal data**, including by highlighting limitations to **surveillance measures which must be targeted**. It further underlined the importance of more transparency being introduced in all processes of content moderation.

The Recommendation referred to the responsibilities of online content providers to respect fundamental rights and freedoms: in that regard, the Recommendation referred to the need for **regular due diligence assessments of their compliance and to provide their products and services without any discrimination**. The transparency and accountability of terms of conditions should be ensured, as well as that content moderation takes place in a transparent and non-discriminatory manner. The Recommendation acknowledged the limited ability of automated means to assess context and called on ensuring human review in all appropriate cases effective

[137] OECD (2020), 'Current Approaches to Terrorist and Violent Extremist Content among the Global Top 50 Online Content-Sharing Services'; Transparency reporting on terrorist and violent extremist content online - An update on the global top 50 content sharing services'.

[138] The reports are based on a questionnaire that covers 12 main topics, is designed to be answerable by services of all sizes, and is intended to produce a baseline level of transparency.

[139] Council of Europe (2018), Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries (Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies).

remedies and dispute resolution systems that provide prompt and direct redress and an impartial and independent review of the alleged violation.

More recently, in May 2021, the Steering Committee for Media and Information Society (CDMSI) published a Guidance Note providing practical guidance to Member States of the CoE, taking into account existing good practices, for policy development, regulation and use of content moderation in the online environment in line with their human rights obligations under the ECHR[140]. The Guidance Note is also addressed to internet intermediaries who have human rights responsibilities of their own. It identifies in particular, and elaborates on, a number of key principles that should guide **a human rights-based approach to content moderation**, such as human rights by default, transparency, clear legal and operational framework, proportionality, safeguards against over-compliance and discrimination, and independent review mechanisms.

II.2. THE EU LEGAL FRAMEWORK: FROM COORDINATION TO HARD LAW OBLIGATIONS

II.2.1. The EU Internet Forum

Initial regulatory steps in online content moderation took place in the context of EU counter-terrorism efforts. The EU has been seeking to play an active role in steering and influencing private practices and decisions on content removal. In December 2015, the Commission established the EU Internet Forum under the European Agenda on Security to address the misuse of the internet for terrorist purposes[141].

The EU Internet Forum includes two main strands of action: the reduction of accessibility to terrorist content online and the increase in the volume of effective alternative narratives online[142]. In 2019, it expanded its scope to encompass the fight against child sexual abuse online. The Forum envisages the participation of 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[143]. Its mission is to provide a collaborative environment for EU governments, the internet industry, and other partners to discuss and address the challenges posed by the presence of malicious and illegal content online. It also explores possible responses against abuse and exploitation of online platforms by terrorists and violent extremists, as well as other malicious actors, including those that groom children for the purpose of sexual abuse and the production and dissemination of child sexual abuse material online.

[140] Council of Europe (2021), op. cit.

[141] European Commission, European Agenda on Security COM(2015) 185.

[142] See https://home-affairs.ec.europa.eu/networks/european-union-internet-forum-euif_en.

[143] European Commission Press release of 3 December 2015, IP/15/6243.

The work of the EU Internet Forum is manifold: it has, for example, agreed on the EU Crisis Protocol following the Christchurch attack in March 2019[144]; it has decided to take steps to tackle violent right-wing extremist groups' presence online by creating a list of such groups, symbols and manifestoes, with the aim to facilitate online content moderation for industry stakeholders; and it has launched the Civil Society Empowerment Programme (CSEP) to support civil society organisations to combat terrorist and extremist propaganda online.

II.2.2. The EU Internet Referral Unit

The EU Internet Forum has been instrumental in the establishment of the EU Internet Referral Unit (IRU), which was set up in response to policy concerns regarding the promotion or glorification of acts of terrorism and violent extremism by Jihadist groups. In 2015 in particular, the Justice and Home Affairs Council mandated Europol to create the EU Internet Referral Unit (IRU)[145]. as part of the wider EU Internet Forum, with a view to reducing the impact of internet content promoting terrorism or violent extremism.

The EU IRU, hosted by Europol's European Counter Terrorism Centre (ECTC), is an operations centre and 'hub of expertise' established in 2016 that draws attention to terrorist and violent extremist content to online service providers and provides support to EU Member States and third parties in the context of internet-related investigations. The EU IRU is a shared database with more than 200 000 hashes, which are unique digital fingerprints of terrorist videos and images removed from online platforms[146]. IRU's mandate has been formalised in the 2016 Europol Regulation[147]. Whereas its work certainly focuses on 'jihadist propaganda', in particular from Al-Qaida and Daesh, as evidenced by yearly reports dedicated to this topic[148], its scope also covers right-wing extremist content online. In addition, it provides support to Europol's European Migrant Smuggling Centre (EMSC) by flagging Internet content used by traffickers to

---

[144] The EU Crisis Protocol, adopted by the Justice and Home Affairs Ministers in October 2019, is a voluntary mechanism that allows EU Member States and online platforms to respond rapidly and in a coordinated manner to the dissemination of terrorist content online in the event of a terrorist attack, while ensuring strong data protection and fundamental rights safeguards.

[145] The European Union Internet Referral Unit at Europol, OPEN ACCESS G'ovT, 2 February 2016, https://www.openaccessgovernment.org/european-unioninternet-referral-unit-europol-2/24158/.

[146] 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 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. 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. European Commission Press release of 6 December 2017, IP/17/5105; European Commission Statement of 5 December 2018, Statement/18/6681; European Commission Press release of 7 October 2019, IP/19/6009.

[147] Regulation (EU) 2016/794 of the European Parliament and of the Council of 11 May 2016 on the European Union Agency for Law Enforcement Cooperation (Europol) and replacing and repealing Council Decisions 2009/371/JHA, 2009/934/JHA, 2009/935/JHA, 2009/936/JHA and 2009/968/JHA [2016] OJ L135/53.

[148] For example, see https://www.europol.europa.eu/cms/sites/default/files/documents/Online_Jihadist_Propaganda_2021_in_review.pdf.

offer smuggling services to migrants and refugees. The EU IRU aims to identify terrorist content online and subsequently map its trace on the internet. It does so by collecting publicly available information about the specific content and creating a 'referral package' to be used for different purposes ranging from assessing the threat to supporting investigations and suggesting referral to online service providers either by Europol or Member States. On the basis of the referrals, the EU IRU coordinates Referral Action Days, which facilitate direct cooperation among law enforcement representatives in EU Member States[149].

In practice, **IRU sends a referral to companies, accompanied by a substantiation of the reasons why specific online content violates their community guidelines.** The legal mechanism of referral gives public authorities the possibility to use private ToS tactically, fulfilling their counterterrorist objective without leveraging classical judicial means. From a private company perspective, the IRU's referral mechanism operates as follows: Europol does not refer content on the basis that it violates the law of an EU Member State, rather based on a violation of the providers' own guidelines. As a result, **companies remain the ones to make the final decision on content removal**[150].

II.2.3. The TERREG Regulation

Both the EU Internet Forum and the EU IRU have increased the voluntary cooperation between government, tech companies, and civil society to counter online terrorism and violent extremism as part of a public-private partnership. However, not all affected hosting service providers have engaged in the Forum and the scale and pace of progress among hosting service providers has not been sufficient to adequately address the problem. As a result, enhanced action had to be taken in the form of adoption of hard law obligations.

Directive (EU) 2017/541 on combating terrorism only briefly touches upon aspects related to online content moderation[151]. The latter requires Member States to undertake the necessary measures to ensure the prompt removal of online content constituting a public provocation to commit a terrorist offence hosted in their territory[152]. Furthermore, the Directive prescribes that Member States must also endeavour to obtain the removal of such content hosted outside their territory. When removal at its sources is not feasible, Member States may take measures to block access to such content towards the internet users within their territory[153]. In addition, those measures must be laid down following transparent procedures and provide adequate safeguards, in particular to ensure that those measures are limited to what is necessary and

[149] Since 2015, the EU IRU has organised a total of 23 Referral Action Days.

[150] Bellanova, R. and de Goede, M. (2022), op. cit.

[151] Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA, OJ [2017] L88/6.

[152] Ibid, Article 21(1).

[153] Ibid, Article 21(2).

proportionate, and that users are informed of the reason for those measures. Safeguards relating to removal or blocking must also include the possibility of judicial redress.

In the transposition of the Directive two main types of measures have been adopted; 'notice-and-takedown' measures and 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[154]. As pointed out, such **measures have differed on several points among the Member States in terms of the offences covered, time limits for removal, and consequences of non-compliance**[155]. None of these measures however have been directed at the online content providers and no EU-wide approach on proactivity on their behalf was foreseen. Overall, as the Commission found, the transposition has been 'uneven'[156].

In view of the growing inclination of terrorist groups to misuse the Internet in order to recruit supporters, facilitate terrorist attacks, and spread fear in the public domain, the **European Commission proposed in 2018 a Regulation on the Removal of Terrorist Content Online (TERREG)**[157], adopted in April 2021, officially published in June 2021, and entered into force on 7 June 2022[158]. The TERREG Regulation aims to prevent the dissemination of terrorist content online by imposing upon online service providers the obligation to remove such content within specific and stringent time limits.

'Terrorist content' is defined in Article 2(7) in a broad manner and includes any material which incites the commission of a terrorist offence[159], which in turn is controversially defined under

[154] European Commission, Report from the Commission to the European Parliament and the Council based on Article 29(1) of Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA, COM(2020) 619 final, pp. 15-16.

[155] Ibid.

[156] Ibid. For example, Greece did not transpose this article. Its legislation only covers the seizing of digital data in the context of criminal investigations, but not removal or blocking of online content. The first sentence of Article 21(1) of the Directive, with regard to removal of online content, does not seem to be transposed by two Member States (Bulgaria and Poland) as their laws only refer to blocking of content. Croatia and Latvia provide measures that may result in the removal of content, but there appears to be no explicit obligation to this end. In the Czech Republic, the law allows the national authorities to request the removal of online content, but it does not provide for the relevant procedure. The second sentence of Article 21(1), which encourages Member States to obtain the removal of content hosted outside their territory, is covered by sixteen Member States. This includes the two Member States (Bulgaria and Poland) that cover solely blocking and thus not removing online content. The option under Article 21(2) in relation to blocking of access to content, when removal at its source is not feasible, is transposed by eighteen Member States. Regarding Article 21(3), in Belgium, Finland, Luxembourg, Poland. and Slovenia the legislation does not seem to require informing users on the reason for content removal.

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

[158] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online [2021] OJ L172/79.

[159] The full definition: 'terrorist content' means one or more of the following types of material, namely material that: (a) incites the commission of one of the offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541, where such material, 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;

EU law by Directive (EU) 2017/541 as a criminal action aimed at seriously intimidating the population, compelling governments to carry out or abstain from certain acts, or seriously destabilising the order of the targeted countries. As a result, terrorist content is intertwined with an overly capacious definition of terrorism which deviates from UN definitions of terrorism (e.g. Security Council Resolution 1566 or the 1999 Terrorism Financing Convention) and from the Council of Europe Convention on the Prevention of Terrorism[160].

According to its Article 3(1), a national 'competent authority […] shall have the power to issue a removal order requiring the hosting service provider to remove terrorist content or disable access to it' (Art. 3(1)). **This mechanism thus obliges platforms to act upon targeted pieces of online content, in striking resemblance to judicial practices**[161]. **The removal mechanism partially disentangles removal processes from strictly judicial ones**. The reference to 'competent authorities' signifies that these are not expected to be only judicial authorities, as Member States may define one or more public agencies that are, in each country, entitled to produce removal orders[162]. These competent authorities are entrusted with a series of tasks: (a) to issue removal orders; (b) scrutinise cross-border removal orders; (c) oversee the implementation of specific measures to protect its services against dissemination to the public of terrorist content; (d) impose penalties for infringements of the Regulation by hosting service providers[163].

Furthermore, **service providers have proactive duties in cases of 'exposure' to terrorist content and on the basis of their own terms and conditions**[164]. To enable the detection and identification and removal of terrorism content, the TERREG Regulation legitimises the use of automated tools if service providers consider this to be appropriate and necessary to effectively address the misuse of their services for the dissemination of terrorist content[165]. The adoption of the TERREG Regulation also has direct impact on the work of Europol in this context; the relationship between the European Counter Terrorism Centre, Member States, and private

(b) solicits a person or a group of persons to commit or contribute to the commission of one of the offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541; (c) solicits a person or a group of persons to participate in the activities of a terrorist group, within the meaning of point (b) of Article 4 of Directive (EU) 2017/541; (d) provides instruction on the making or use of explosives, firearms or other weapons or noxious or hazardous substances, or on other specific methods or techniques for the purpose of committing or contributing to the commission of one of the terrorist offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541; (e) constitutes a threat to commit one of the offences referred to in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541.

[160] For criticism see among others Karaliota, N., Kompatsiari, E., Lampakis, C. and Kaiafa-Gbandi, M. (2020), *The New EU Counter-Terrorism Offences and the Complementary Mechanism of Controlling Terrorist Financing as Challenges for the Rule of Law*, Brill; Gherabaoui,T. and Scheinin, M. (2021), 'Time to Rewrite the EU Directive on Combating Terrorism', Verfassungsblog, 25 January 2022, https://verfassungsblog.de/time-to-rewrite-the-eu-directive-on-combating-terrorism/.

[161] Bellanova, R. and de Goede, M. (2022), op. cit.

[162] Recital 35. See Bellanova, R. and de Goede, M. (2022), op. cit.

[163] Article 12.

[164] Article 5.

[165] Ibid. Recital 25.

providers will change, as well as the responsibility of the IRU. **Member States will be able to request the removal of content from online service providers through a platform called PERCI, built by Europol and managed by the IRU**.

Moreover, the removal of terrorist content or disabling access to it following a request must be done as soon as possible and in any event within one hour of receipt of the removal order[166]. **This very tight deadline, which may incentivise the providers to use automated content moderation tools** in order to identify and delete terrorist content, has been a key criticism leading 61 human rights organisations to send a joint letter to the European Parliament asking its members to vote against the proposal[167].

II.2.4. The Recast Europol Regulation

During the final stages of the negotiations on the TERREG Regulation, in December 2020, the Commission adopted a proposal for **the recast of the Europol Regulation** aiming to enable the agency to step up its support to Member States in fighting serious crime and terrorism and tackling emerging security threats[168]. In February 2022, co-legislators reached a provisional agreement on the proposal from 28 June 2022, the amended Europol Regulation enters into force and becomes applicable[169].

Among its many reforms, the amended Europol Regulation **expands the scope of law enforcement authorities cooperation with service providers in the fight against crime and terrorism, including in the field of digital surveillance of online content**[170]. In particular, Article 26a prescribes that **Europol will be able to receive personal data directly from private parties** and analyse that data to identify those Member States that could open investigations into related crimes.

Separate provisions have been added regarding the role of Europol in supporting Member States to prevent the dissemination of online content related to terrorism and violent extremism as conforms with the TERREG. Europol will provide the support necessary for competent authorities of the Member States to interact with private parties, in particular by providing the necessary infrastructure for such interaction, for example, when competent

---

[166] Article 3(3).

[167] Joint Letter to EU Parliament: Vote Against Proposed Terrorist Content Online Regulation, 25 March 2021 https://www.hrw.org/news/2021/03/25/joint-letter-eu-parliament-vote-against-proposed-terrorist-content-online,

COM(2020) 796, 9 December 2020.

[169] Regulation (EU) 2022/991 of the European Parliament and of the Council of 8 June 2022 amending Regulation (EU) 2016/794, as regards Europol's cooperation with private parties, the processing of personal data by Europol in support of criminal investigations, and Europol's role in research and innovation [2022] OJ L169/1.

[170] Article 1, Article 18a.

authorities of the Member States refer terrorist content online, send removal orders concerning such content to online service providers pursuant to the TERREG[171].

Furthermore, **Europol can support Member States' actions to address dissemination of terrorist content in the context of 'online crisis situations[172]'** stemming from ongoing or recent real-world events, online dissemination of child sexual abuse material and to support the actions of online service providers in compliance with their obligations under EU law or in acting voluntarily. As a result, it can exchange relevant personal data, including hashes, IP addresses, or URLs related to such content with private parties under certain conditions[173].

On the 27 June 2022, **the European Data Protection Supervisor (EDPS) regretted that the newly expanded Europol's mandate has not been accompanied by strong data protection safeguards allowing for an effective oversight or supervision of the Agency**[174]. The EDPS expressed serious concerns about Europol's new competence to process the personal data of individuals with no established link to criminal activity and the retroactive authorisation for Europol to process large sets of data shared by Member States before the adoption of the new Regulation.

II.2.5. Fighting child sexual abuse online

As has been made evident in this Report thus far, in addition to terrorist-related content, **another key area in which enhanced cooperation concerning online content moderation has been taking place is that of child sexual abuse**. In its EU Strategy for a More Effective Fight Against Child Sexual Abuse of June 2020[175], the Commission set out a comprehensive response to the growing threat of child sexual abuse both offline and online, by improving prevention, investigation, and assistance to victims. Among its eight initiatives the Strategy referred to the role of industry to ensure the protection of children when using the services they offer, and improve protection of children globally through multi-stakeholder cooperation.

On 24 March 2021, the European Commission adopted its comprehensive EU Strategy on the Rights of the Child, which proposes reinforced measures to protect children against all forms of violence, including online abuse[176]. In addition, it invited companies to continue their efforts to detect, report and remove illegal online content, including online child sexual abuse, from their platforms and services.

---

[171] Recital 42 and Article 4(m).

[172] As defined in Article 2(t).

[173] Recital 43 and Article 4(x). Such exchanges of personal data should only take place for the purposes of removing terrorist content and online child sexual abuse material, in particular where the exponential multiplication and virality of that content and material across multiple online service providers are anticipated.

[174] European Data Protection Supervisor (EDPS), Amended Europol Regulation Weakens Data Protection Supervision, Press Release, 27 June 2022, Amended Europol Regulation weakens data protection supervision | European Data Protection Supervisor (europa.eu).

[175] European Commission, EU strategy for a more effective fight against child sexual abuse, COM(2020) 607, 24 July 2020, p. 2.

[176] See https://ec.europa.eu/info/sites/default/files/ds0821040enn_002.pdf.

Whereas certain providers already voluntarily use technologies to detect, report, and remove online child sexual abuse on their services, the Commission has deemed the action insufficient to address the misuse of online services[177]. Furthermore, while Member States have started preparing and adopting national rules to fight against online child sexual abuse, approaches vary considerably[178].

As a result, in May 2022, **the Commission proposed a Regulation on combating child sexual abuse**. These rules will complement those regarding the criminalisation of sexual abuse and sexual exploitation of children and child sexual abuse materials by the Child Sexual Abuse Directive, adopted in 2011[179]. The proposal builds on the so-called Interim Regulation on combating online child sexual abuse, which is restricted to the voluntary actions of a limited number of online services and applicable for a period of maximum three years[180].

In a nutshell, under the proposed rules **providers will be obliged to detect, report, and remove or disable child sexual abuse material on their services**. This will not only concern child sexual abuse material (which has been verified as such by authorities), but providers must also proactively search for new photos and videos, as well as evidence of text-based 'grooming', which will require use of AI-based tools and techniques to scan private conversations. They will have to assess and mitigate the risk of misuse of their services and the measures taken must be proportionate to that risk and subject to robust conditions and safeguards[181]. The proposal further lays down obligations on providers to disable access to Child Sexual Abuse Material (CSAM). Furthermore, a new independent EU Centre on Child Sexual Abuse (EU Centre) will facilitate the efforts of service providers by acting as a hub of expertise[182].

The proposal has already received fierce criticism. In a joint letter signed by numerous civil society and professional (trade union) organisations, **the proposal has been condemned for embracing scanning and surveillance technologies for identifying illegal content and for transforming the internet into a space that will be 'dangerous for everyone's privacy, security,**

[177] The measures taken by providers vary widely, with the vast majority of reports coming from a handful of providers, and a significant number take no action. The quality and relevance of reports received by EU law enforcement authorities from providers also varies considerably. See European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse COM(2022) 209, 11 May 2022.

[178] Ibid, p. 2.

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

[180] The Interim Regulation lays down temporary limited rules to enable providers of certain communication services to use technologies for the detection, reporting and removal of online child sexual abuse on their services, thereby derogating from certain obligations laid down in Directive 2002/58/EC (ePrivacy Directive). See Regulation (EU) 2021/1232 of the European Parliament and of the Council of 14 July 2021 on a temporary derogation from certain provisions of Directive 2002/58/EC as regards the use of technologies by providers of number-independent interpersonal communications services for the processing of personal and other data for the purpose of combating online child sexual abuse (Text with EEA relevance) [2021] OJ L 274/41.

[181] Chapter II.

[182] Chapter IV.

and free expression'[183]. In making providers liable for the private messages shared by their users, companies will be forced to utilise risky and inaccurate tools to manage online content. It may discourage child abuse survivors from coming forward or whistle-blowers if they know that confiding to a trusted person via a private message could mean that the provider could flag the message for review. Secure messenger service providers such as Signal will be required to alter their services, thus creating risks for anyone relying on them.

Similarly, the EDPS and the European Data Protection Board (EDPB) issued a Joint Opinion on 28 July 2022 in which they underlined the lack of proportionality and necessity of the envisaged detection measures in the proposal[184]. They highlighted that **the intrusiveness of a proposal granting the possibility to access the content of communications on a generalized and indiscriminate basis, and its serious interference with the essence of the rights of privacy and data protection**. The EDPS and the EDPB recommended that the measures included in the Proposal dealing with the detection of solicitation of children and grooming should be deleted.

II.3. THE DIGITAL SERVICES ACT

II.3.1. Legal background

In parallel to the aforementioned steps taken in the law enforcement context, **a number of EU soft and hard law initiatives have emerged without a sector-specific approach**. Compared to the previous steps that have largely taken place in the past few years, the **EU regulatory framework on content moderation has been rather old, complex, and scarce**.

Before the adoption of the proposal for a Digital Services Act (DSA), there existed some horizontal rules applicable to all categories of online platforms and to all types of content in accordance with the e-Commerce Directive[185]; certain stricter rules applicable to Video-Sharing Platforms (VSPs) and to certain types of illegal content online, for example the revised Audio-visual Media Service Directive (AVMSD)[186]. Those rules were complemented by the vertical rules applicable to the four types of illegal content, as discussed in the previous Sections of this Report, as well as several codes of practices have been adopted by the main online platforms to better tackle illegal and harmful content, as noted in *Section I* above.

---

[183] See https://edri.org/our-work/protecting-digital-rights-and-freedoms-in-the-legislation-to-effectively-tackle-child-abuse/. For a first analysis see Quintel, T. (2022), 'The Commission Proposal on Combatting Child Sexual Abu–e - Confidentiality of Communications at Risk?', *European Data Protection Law Review*, Vol. 8 No 2, p. 282.

[184] European Data Protection Supervisor and European Data Protection Board (2022), Joint Opinion 4/2022, on the Proposal for a Regulation laying down rules to prevent and combat child sexual abuse, 28 July 2022, 22-07-28_edpb-edps-joint-opinion-csam_en.pdf (europa.eu).

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

[186] Directive 2010/13/EU 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) (Text with EEA relevance) [2010] OJ L 95/1.

The **e-Commerce Directive** in particular was adopted in 2000, when online platforms were in their infancy and many current technologies and applications did not exist yet. Its overarching aim has been to stimulate cross-border trade by removing legal obstacles to the exercise of the freedom of establishment and the freedom to provide services stemming from divergences in legislation, legal uncertainty as to which national rules apply, and the extent to which Member States may control services originating from another Member State[187].

In relation to content moderation, Article 14 of the e-Commerce Directive provides exemptions from the national liability regime for three categories of online platforms, i.e., mere conduit, caching, and hosting subject to requirements[188]. These exemptions are horizontal, which means that many types of illegal content are covered as well as criminal and civil liability. Article 14 prescribes that 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'. In case the platform has such knowledge or awareness, it can nonetheless benefit from the liability exemption if it 'acts expeditiously to remove or to disable access to the information'. The concepts 'actual knowledge' and 'acting expeditiously' have given rise to debate as to their meaning[189].

Another central provision is Article 15, a general monitoring prohibition, according to which 'Member States shall not impose a general obligation on providers (…) to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity'. Nevertheless, online platforms could decide, on a voluntary basis, to carry out spot checks on online content. Though permitted, it may signify that the online provider could be considered as playing an active role, and thus not fall within the liability exception[190].

In terms of duty of cooperation with the competent authorities, Article 15(2) of the Directive requires providers promptly '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'. Finally, and in line with the efforts discussed in *Section I*, Article 16 of the Directive encourages co- and self-regulation to implement the rules of the Directive.

---

[187] Recital 5.

[188] For an interpretation see Cases C-236/08 to C-238/08 Google France v Louis Vuitton EU:C:2010:159 introducing the requirement of neutrality, which state that t: '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'.

[189] A. De Streel, et al. (2020), Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform, Study requested by the IMCO Committee, p. 21.

[190] Ibid, p. 22.

The Directive has been subject to three evaluations, with the latest one from 2016 noting the increasing importance of online platforms and the risks for a fragmented Digital Single Market[191]. A strategy emerged in this respect comprising of three strands: a) adapting sectoral hard law; (b) giving more guidance on the interpretation of the less clear provisions of the e-commerce Directive, in particular regarding the Notice-and-Takedown and the reliance on voluntary preventive measures; and (c) encouraging coordinated EU-wide co and self-regulation for the illegal materials which are particularly harmful.

Following that evaluation, in 2016, **service providers agreed on a Code of Conduct** with the EU, aiming 'to have in place clear and effective processes to review notifications regarding illegal hate speech ... [and] to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content'[192]. Facebook, Twitter, YouTube, and Microsoft agreed to adhere to the Code. Snapchat, Instagram, Dailymotion, Google+, and Jeuxvideo also subsequently agreed to adhere to the Code. At that time, Commissioner Jourova resisted the idea of adopting legislation, leaving, however, the matter open in case companies do not self-regulate satisfactorily by May 2018[193].

In the meantime, in September 2017, **the Commission released guidelines and principles requesting that online platforms increase the proactive prevention, detection, and removal of illegal content**. It do so under a remarkably expansive approach that included not only material inciting terrorism, illegal hate speech, or child sexual abuse, but also content relating to 'trafficking in human beings [,] … violations of intellectual property rights, product safety rules, illegal commercial practices online, or online activities of a defamatory nature'[194].

On 1 March 2018, the Commission adopted a **Recommendation on measures to effectively tackle illegal content online**[195], building upon a Communication of September 2017[196]. The Recommendation included a specific chapter laying down numerous measures to effectively stem the uploading and sharing of terrorist propaganda online, such as improvements to the referral process, am one-hour timeframe for responding to referrals, more proactive detection, effective removal, and sufficient safeguards to accurately assess terrorist content.

---

[191] Commission Staff Working Document of 25 May 2016, Online Platforms, SWD(2016) 172.

[192] European Commission on Code of Conduct on Countering Illegal Hate Speech Online, May 31 2016, http://ec.europa.eu/justice/fundamentalrights/files/hate-speech-codeof_conduct-en.pdf.

[193] Boffey, D., 'EU justice commissioner resists calls for legislation on online hate speech', *The Guardian*, 28 September 2017), https://www.theguardian.com/uknews/2017/sep/28/eu-justice-commissioner-resists-calls-for-legislation-on-onlinehate-speech.

[194] Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms, at 2, 6, COM (2017) 555 final, 28 September 28 2017.

[195] European Commission, Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online [2018] OJ L 63/50.

[196] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regio–s - Tackling Illegal Content Online Towards an enhanced responsibility of online platforms, COM(2017) 555, 28 September 2017.

While the e-Commerce Directive has been the cornerstone of the Internal Market for the last twenty years, Members of the European Parliament's Internal Market and Consumer Protection Committee (IMCO) noted that the digital single market is affected by increasing fragmentation in tackling illegal content online, difficulties regarding market entry, concerns over consumer welfare, and ineffectiveness of enforcement and cooperation between Member States[197].

As a result, and in anticipation of a new Digital Services Act (DSA), which was proposed in the Commission's 2020 Work Programme, **the IMCO committee took the initiative to prepare a legislative own-initiative report with recommendations on a DSA** that would improve the functioning of the single market (Rapporteur: MEP Alex Agius Saliba)[198]. According to the IMCO report, the DSA should work along two pillars. Pillar one, on the one hand, was set out to ensure trust and safety online by increasing responsibilities, obligations, and liabilities for digital services. Pillar two, on the other hand, should bring *ex ante* regulation for big platforms, so-called 'gate-keepers'. These *ex ante* measures aimed at preventing market failures caused by gatekeepers' anti-competitive behaviour. Next to IMCO's report, the Legal Affairs (JURI) Committee similarly drafted a legislative own-initiative report on the DSA[199], while the Civil Liberties Justice and Home Affairs (LIBE) committee issued a non-legislative own-initiative report on the same topic[200].

The Commission incorporated much of what was proposed by the IMCO committee and split the reform of the digital single market into two legislative packages that were reflective of the two pillar-logic and officially proposed on 15 December 2020 **the Digital Services Act (DSA)**[201] and **the Digital Markets Act (DMA)**[202]. Together they constitute a set of rules applicable across the EU with the aim of creating a safer digital space and establishing a level playing field to foster innovation, growth, and competitiveness, both in the European Single Market and globally.

On 23 April 2022 and following rather swift negotiations, **the European Parliament and the Council reached political agreement on the DSA.** At the time of writing, the final text of the DSA has not been published in the *Official Journal of the EU*. The following paragraphs provide a

[197] European Parliament, Report with recommendations to the Commission on the Digital Services Act: Improving the functioning of the Single Market, P9_TA(2020)0272.

[198] Ibid.

[199] European Parliament, Draft Report with recommendations to the Commission on a Digital Services Act: adapting commercial and civil law rules for commercial entities operating online (2020/2019(INL)), PE650.529v01-00.

[200] European Parliament, Draft Report on the Digital Services Act and fundamental rights issues posed (2020/2022(INI)), PE650.509v01-00.

[201] European Commission, Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020) 825, 15 December 2020.

[202] European Commission, Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act), COM(2020) 842, 15 December 2020.

concise outline of the relevant rules in line with the adopted text published by the European Parliament dated from 5 July 2022, which reflects the agreement between the co-legislators[203]. For the full assessment, also cross-refer with Section III.2 below.

II.3.2. The DSA explained

The DSA lays down EU-wide due diligence obligations that will apply to all digital services that connect consumers to goods, services, or content depending on their roles, size, and impact on the online ecosystem[204]. A key component of these duties concerns **new procedures for fast removal of illegal content**.

Its scope of application concerns various groups of online intermediary services, which have their place of establishment or are located in the EU, irrespective of the place of establishment of the providers of those services[205]. Thus, the obligations will apply in the EU single market without discrimination. According to Article 2(f), these **online intermediary services** involve:

> (a) a 'mere conduit' service, that consists of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network, such as Internet access providers, domain name registrars;
>
> (b) a 'caching' service, that consists of the transmission in a communication network of information provided by a recipient of the service, involving the automatic intermediate and temporary storage of that information, performed for the sole purpose of making more efficient the information's onward transmission to other recipients upon their request and
>
> (c) a 'hosting' service that consists of the storage of information provided by, and at the request of, a recipient of the service. For very large online platforms, reaching more than 10 % of 450 million consumers in Europe, that pose particular risks in the dissemination of illegal content and societal harms specific rules are also foreseen.

'Illegal content' is defined in the DSA as 'any information, which in itself or in relation to an activity including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law'[206].

Article 5 of the DSA has maintained the general rule according to which providers of hosting services are not liable for user-generated content unless they have actual knowledge about illegal online activity or upon obtaining such knowledge or awareness, act expeditiously to

---

[203] For the agreed text see https://www.europarl.europa.eu/doceo/document/TA-9-2022-0269_EN.html. For the adopted text see https://www.europarl.europa.eu/doceo/document/TA-9-2022-0269_EN.html.

[204] Article 1.

[205] Article 1a.

[206] Article 2(g).

remove or to disable access to the illegal content[207]. Providers can be excluded from liability because they, in good faith and in a diligent manner, carry out voluntary own-initiative investigations or take other measures aimed at detecting, identifying, or disabling of access to illegal content, or take measures to comply with EU or national law[208]. Furthermore, the DSA does not impose a general obligation for service providers to monitor the information transmitted or stored or to actively seek facts or circumstances indicating illegal activity[209].

The following sub-sections provide an overview of the DSA by looking into the duties to act against illegal content and to provide information; the due diligence duties to hosting services; the rules specifically oriented at very large service providers; and the provisions on transparency and supervision.

II.3.2.a. Duties to act against illegal content and to provide information

With regard to orders to act against illegal content, Article 8 requires providers to inform the issuing authority, which may be a judicial or administrative one, or any other authority specified in the order of any follow-up given to the orders, without undue delay about whether and when the order was applied[210].

Providers may also be asked to provide specific information about one or more specific individual users[211]. Similarly, providers will then have to inform the issuing authority or any other authority specified in the order of the effect given to the order, specifying if and when the order was applied.

The DSA requires Member States to establish Digital Services Coordinators in each Member State; in cases of orders against 'illegal content' or to provide information the Digital Services Coordinator from the Member State of the issuing judicial or administrative authority (or specified in the order) must receive the order and transmit a copy to all Digital Services Coordinators[212]. In addition, providers must inform users (at the latest at the time when the order was applied) of the order received and the effect given to it, or where applicable, by the time provided by the issuing authority in its order.[213]

The European Parliament attempted to introduce explicit rules for users on effective remedies against orders, including restoration of the content that has been erroneously considered as

[207] Article 5(1).

[208] Article 6.

[209] Article 7.

[210] This obligation is subject to a series of requirements, for example on the elements of the order, which are set out in Article 8.

[211] Article 9.

[212] Article 8(2d) and 8(3); Article 9(2a) and 9(3).

[213] Article 8(3a); Article 9(3a). Such information must at least include the statement of reasons and the redress possibilities. In relation to orders to act against illegal content the information must also include the territorial scope of the order.

illegal beyond the remedies under EU data protection law[214]. However, reference to the possibility of restoration has only remained in the Preamble[215].

II.3.2.b Due diligence duties to hosting services

In addition to requirements for providers to comply with national orders, **the DSA imposes new mechanisms allowing users, both persons and legal entities, to flag illegal content online**. Article 14 states that those mechanisms shall be 'easy to access, user-friendly, and allow for the submission of notices exclusively by electronic means'. The submission of a notice including these elements will be considered to give rise to actual knowledge or awareness for the purposes of Article 5, where it allows a diligent provider of hosting services to identify the illegality of the relevant activity or information without a detailed legal examination[216].

Where contact information would be provided by the individual or entity that submitted the request then a confirmation of receipt of the notice will be given without undue delay[217]. In addition, the provider shall also, without undue delay, notify that individual or entity of its decision in respect of the information to which the notice relates, providing information on the redress possibilities in respect of that decision[218]. According to Recital 40, the notification mechanism should allow, but not require, the identification of the notice provider. However, for some types of items of information notified, the identity of the notice provider might be necessary to determine whether it constitutes illegal content, as alleged. All received notices must be processed 'in a timely, diligent, non-arbitrary and objective manner'[219]. In cases of use of automated means for processing or decision-making, information on such use must be included in the notification to the individual or entity[220].

Providers must provide a clear and specific statement of reasons to the affected users regarding any restrictions imposed, which involve: any restrictions of the visibility of specific items of information provided by the recipient of the service, including removal of content, disabling access to content, or demoting content; suspension, termination, or other restriction of monetary payments (monetisation) suspension or termination of the provision of the service in whole or in part; or suspension or termination of the recipient's accounts[221]. This requirement will only apply where the relevant electronic contact details are known to the provider. It will apply, at the latest, when the restriction is imposed, and regardless of why or how it was imposed. In turn, this obligation shall not apply where the information is deceptive

[214] See agreed text, op. cit,, p. 280.

[215] Recital 33a.

[216] Article 14(3).

[217] Article 14(4).

[218] Article 14(5).

[219] Article 14(6).

[220] Ibid.

[221] Article 15(1).

high-volume commercial content[222]. However, these requirements will not apply in respect of any orders issued under Article 8, in which case the prescriptions of Article 8(3a), as stated above, are applicable.

Article 15a introduces additional duties in cases of suspected criminal offences. Providers becoming aware of any information giving rise to a suspicion that a criminal offence involving a threat to the life or safety of persons has taken place, is taking place, or will take place, must inform the law enforcement or judicial authorities of the Member State(s) concerned. If the latter cannot be identified with reasonable certainty then the information must reach the authorities of the Member State in which it is established or has a legal representative, Europol, or both[223].

Additional rules for online platforms in particular are also foreseen, but these are not applicable to micro and small enterprises[224].

Article 17 requires providers of online platforms to provide users including those who submit a notice, for a period of at least six months[225] the **access to an effective internal complaint-handling system**. The system must enable the complaints to be lodged electronically and free of charge, against the decision taken by the provider of the online platform upon the receipt of a notice or against **the following decisions on the ground that the information provided is illegal content or incompatible with its terms and conditions**: (a) decisions whether or not to remove or disable access to or restrict visibility of the information; (b) decisions whether or not to suspend or terminate the provision of the service, in whole or in part, to the users; (c) decisions whether or not to suspend or terminate the recipients' account; and (d) decisions whether or not to suspend, terminate or otherwise restrict the ability to monetize content provided by the recipients.

The online complaint mechanism must be easy to access, user-friendly and enable and facilitate the submission of sufficiently precise and adequately substantiated complaints[226]. The complaints must be handled in a timely, non-discriminatory, diligent, and non-arbitrary manner[227]. However, **a ten-day deadline for processing complaints, as proposed by the European Parliament has not been included in the final text.** Where a complaint contains sufficient grounds for the provider to consider that its decision not to act upon the notice is unfounded, or that the information to which the complaint relates is not illegal and is not incompatible with its terms and conditions, or contains information indicating that the complainant's conduct does not warrant the measure taken, it must reverse its decision

[222] Article 15(1a).

[223] Article 15a(2).

[224] On the meaning of this see Annex to Recommendation 2003/361/EC.

[225] Starting from the day they are informed about the decision.

[226] Article 17(2).

[227] Article 17(3).

without undue delay[228]. The complainants must be informed accordingly about this decision. **Any decisions on the complaints must not be taken solely by automated means**— in that regard and at the behest of the Parliament, it has been added that **decisions must be taken 'under the control of appropriately qualified staff'**[229].

An out-of-court dispute settlement mechanism is also foreseen, which may even be established by the Member States[230]. According to Article 18, all users must be entitled to select any certified out-of-court dispute settlement body for the resolution of disputes relating to those decisions. This would not prevent users to initiate, at any stage, proceedings to contest those decisions by the providers of online platforms before a court in accordance with the applicable law[231].

The certified out-of-court dispute settlement body shall not have the power to impose a binding solution on the parties[232]. Certification will be done by the Digital Services Coordinator for a (renewable) maximum period of five years if the system complies certain conditions. To ensure impartiality of the persons charge of dispute resolution, the Parliament had suggested including an additional requirement; that **they must commit not to work for the online platform or a professional organisation or business association of which the online platform is a member** for a period of three years after their position in the body has ended, and have not worked for such an organisation for two years prior to taking up this role.

Finally, **online platforms must take measures to enable trusted flaggers, acting within the designated area of expertise, to submit notices which must be given priority**[233]. For an entity to be certified as a trusted flagger, the Digital Services Coordinator must assess the following conditions: that it has particular expertise and competence in detecting, identifying, and notifying illegal content, independence from any provider and diligence, accuracy, and objectivity in carrying out its activities.

II.3.2.c. Providers of very large online platforms

Section 4 of the DSA contains **additional duties addressed to very large online platforms**, on conducting risk assessments, taking mitigation measures, being subject to a crisis response mechanisms, annual audits, and establishing a compliance function. In particular, according to Article 26 they must carry out **risk assessments on a yearly basis**, as well as every time they wish to deploy functionalities that are likely to have a critical impact[234].

---

[228] Ibid.

[229] Article 17(5). In that regard, a Recital will also mention at the behest of the Parliament that recipients of the service are given the possibility, where necessary, to contact a human interlocutor at the time of the submission of the complaint.

[230] Article 18(4).

[231] 18(2).

[232] Article 18(1)

[233] Article 19.

[234] Article 26.

In connection to any risks identified, **providers must put in place reasonable, proportionate, and effective mitigation measures**[235]. In cases of a 'crisis', defined as extraordinary circumstances leading to a serious threat to public security or public health, Article 27a prescribes that **the Commission** upon a recommendation by a Board **may require very large online platforms or search engines to take certain actions**[236]. In addition, very large online platforms must be subject to yearly audits for the purposes of which they must cooperate with the organisations conducting them in an effective, efficient, and timely manner[237]. These organisations must be independent and with proven expertise, objectivity, and professional ethics[238]. These providers must also establish a compliance function, independent of operational functions, to monitor compliance with the DSA[239].

II.3.2.d Transparency and supervision: Zooming into the role of the Digital Services Coordinators and the Commission

The DSA imposes transparency reporting duties to all providers[240]**, additional ones to online platforms in general**[241] **and some more to very large online platforms**[242]. Supervision of providers and enforcement is entrusted to one or more competent authorities designated at the national level, one of which shall act as the Digital Services Coordinator[243]. As mentioned above, the latter will be bestowed with numerous supervisory and enforcement tasks and must cooperate with national competent authorities, the Commission, and the Board[244].

In the case of very large platforms and search engines **supervision is bestowed to the Commission**, which in order to manage the additional tasks, an annual fee to the providers will be charged[245]. The Commission will also develop expertise and capabilities, including through secondments of Member States' personnel and coordinate the assessment of systemic and emerging issues[246]. Among its tasks the Commission may exercise investigatory powers[247], initiate proceedings against providers infringing the DSA[248], and request for information[249]. In

[235] Article 27.

[236] For example, to assess whether and, if so, to what extent and how the functioning and use of their services significantly contribute to, or is likely to significantly contribute, to a serious threat.

[237] Article 28.

[238] Article 28(2).

[239] Article 32.

[240] Article 13.

[241] Article 23

[242] Article 33.

[243] Article 38(1). For the requirements for the Digital Services Coordinator see Article 39.

[244] Article 38(2).

[245] Article 33b.

[246] Article 49a.

[247] Article 50.

[248] Article 51.

[249] Article 52.

addition, the Commission may conduct interviews and take statements[250], conduct inspections[251], and monitor the effective implementation of the DSA, for example by order providers to give access to databases and algorithms[252]. In cases of non-compliance, the Commission will adopt relevant decisions[253], possibly accompanied by a fine[254].

As regards additional powers, Article 41 prescribes that the Digital Services Coordinators will have investigatory powers to require providers to provide information relating to infringement of the DSA (e.g., retrieving evidence), to carry out or request judicial authorities inspections of any premises to examine, seize, take, or obtain copies of information relating to a suspected infringement, and to ask staff or a representative of the provider for explanations. As for enforcement powers, among others the Digital Services Coordinators will have the power to impose fines or periodic penalty payments[255]. They will also receive complaints by users against providers[256] and draw up activity reports[257]. The Digital Services Coordinators and the Commission must cooperate closely and provide mutual assistance[258] and with one another[259]. They can also launch and lead joint investigations with the participation of other Digital Services Coordinators[260].

An independent group composed of Digital Services Coordinators and chaired by the Commission, named the European Board for Digital Services, is also established to supervise providers[261]. Its tasks will include to support the coordination of joint investigations, support in the analysis of reports and results of audits of very large online platforms or search engines, issue opinions, recommendations or advice the Digital Services Coordinators or the Commission in deciding to initiate proceedings[262]. Finally, the Commission and the Board will foster the drawing up of voluntary codes of conduct[263].

[250] Article 53.

[251] Article 54.

[252] Article 57.

[253] Article 58.

[254] Article 59.

[255] Article 41(2).

[256] Article 43. The responsible Digital Services Coordinator will be determined on the basis of the location or establishment of the user.

[257] Article 44.

[258] Article 44b.

[259] Article 45.

[260] Article 46.

[261] Articles 47-48.

[262] Article 49.

[263] Article 35.

II.4. THE REGULATORY FRAMEWORK IN THE UK

The UK has a pioneering role in online content moderation initiatives. Internet Referral Units (IRUs) have been established since 2010 and their model has been copied in other countries such as France, Belgium, and the Netherlands[264]. Furthermore, the EU IRU was also established with the aid of the UK. The UK's role as a global actor is important: with regard to counterterrorism, it has been disseminating information about its Counter Terrorism IRU (CTIRU) with the aim to coordinate an international response to terrorism propaganda[265]. During the 2017 G7 summit, the UK pressured companies into forming a Global Internet Forum to develop and share new technology and tools to automatically identify and remove content promoting incitement to violence. The UK has also succeeded in obtaining supportive statements by the Five Eyes (Australia, Canada, New Zealand, the United Kingdom, and the United States), in a G20 Summit declaration, and at the European Council meeting.

At the domestic level, **the Terrorist Act 2006** has been used to define terrorist content, including whether the content seeks to encourage terrorism and the dissemination of terrorist material[266]. This change strengthened the existing offence, so that it applies to material that is viewed or streamed online. In October 2017 the Internet Safety Strategy Green Paper was released, which provided a vision for a strategic and coordinated approach to online safety and discussed potential actions to address a range of online harms including harassment, trolling, cyberbullying, sexting, and online abuse.

In April 2019, an **Online Harms White Paper** was published arguing that existing regulatory and voluntary initiatives had 'not gone far or fat enough' to keep users safe[267]. The Paper suggested a **single regulatory framework to tackle various harms** with a duty of care imposed on providers lying at the heart of it, overseen and enforced by an independent regulator. The response to the White Paper was mixed, as some commentators were concerned that **the harms were insufficiently defined and that the approach could jeopardise freedom of expression**[268]. In December 2020, following a period for consultation, a Government response was published confirming that an Online Safety Bill would be introduced imposing duties on online platforms regulated by Ofcom[269]. In the aftermath of the torrent of racist abuse directed towards Black

---

[264] https://hrlr.law.columbia.edu/files/2018/07/BrianChangFromInternetRef.pdf.

[265] Baroness Shields opening speech at the Global Counter Terrorism Form, Gov.UK, 25 January 2017, https://www.gov.uk/government/speeches/baronessshields-opening-speech-at-the-global-counterterrorism-forum.

[266] Furthermore, the Counter-Terrorism and Border Security Act 2019, so that individuals who view terrorist content online could face up to 15 years in prison.

[267] HM Government, Online Harms Paper, April 2019.

[268] Open Rights Group, 'Online Harms: Freedom of expression remains under threat, 'https://www.openrightsgroup.org/blog/online-harms-freedom-of-expression-remains-under-threat/.

[269] Online Harms White Paper: Full government response to the consultation https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response.

English football players following the UEFA Euro 2021 final, calls have increased in the UK to implement new online legislation.

A draft Online Safety Bill was published in May 2021[270]. It is a complex piece of legislation composed of 213 pages, plus 126 pages of Explanatory Notes. The publication of the Online Safety Bill coincided with the negotiations of the DSA at EU level. This has resulted in discussions about the influence of the approaches, with some arguing that in view of the UK's status outside of the EU, it will mean that the DSA will be more influential globally[271]. At the same time, it has been reported that in the post-Brexit era a somewhat 'uneasy and distant relationship' has emerged between the two regulators, resulting in the lack of cooperation on this matter, despite the undoubtedly influential role of the EU, as presented in the previous paragraphs[272].

The Online Safety Bill pushes for more obligations on so-called 'regulated services' — that is 'user-to-user services' and 'search services'[273] that have 'links' with the UK[274] — with regard to three types of content: (a) illegal content; (b) content that is harmful to children; and (c) content that is legal but harmful to adults. The regulatory independence of the UK has resulted in the latter going a step further than the EU requiring the largest platforms to protect users from content that may be lawful, but that could cause physical or psychological damage, such as posts promoting self-harm or disinformation. The Bill further requires regulated services to perform risk assessments and to adopt mitigation measures ('safety duties').

A definition of illegal content is provided in Clause 41(3): 'content consisting of certain words, images, speech or sounds amounts to a relevant offence if the provider of the service has reasonable grounds to believe that: (a) the use of the words, images, speech or sounds amounts to a relevant offence, (b) the use of the words, images, speech or sounds, when taken together with other regulated content present on the service, amounts to a relevant offence, or (c) the dissemination of the content constitutes a relevant offence'. Relevant offences are the following: terrorism offences; child sexual exploitation and abuse offences; an offence set out in secondary legislation; and other offences directed at an individual as the victim[275].

In turn, 'harmful content', such as self-harm or eating disorder content, is defined in Clauses 45 to 47 of the draft Bill. Content will be considered 'harmful' in three circumstances; (a) if designated in secondary legislation as 'primary priority content' that is harmful to children or

[270] A Bill to make provision for and in connection with the regulation by OFCOM of certain internet services; for and in connection with communications offences; and for connected purposes (Bill 004 2022-23). https://bills.parliament.uk/bills/3137/publications.

[271] https://www.politico.eu/article/eu-digital-services-act-uk-online-safety-bill-content-moderation-safety.

[272] Ibid.

[273] 'User-to-user service' means an internet service by means of which content that is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, may be encountered by another user, or other users, of the service. The latter term refers to internet service that is, or includes, a search engine.

[274] Clauses 2 and 3.

[275] Clause 41(4).

'priority content' that is harmful to children or adults; (b) if a service provider has 'reasonable grounds to believe that the nature of the content is such that there is a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact' on a child or adult of 'ordinary sensibilities'; or (c) if a service provider has 'reasonable grounds to believe that there is a material risk' of the dissemination of the content 'having a significant adverse physical or psychological impact' on a child or adult of ordinary sensibilities.

As with the DSA, **the specific scope of the duties varies significantly depending on the nature of the service and the nature of the content**. First, all user-to-user services would have to conduct an illegal content risk assessment[276]; take proportionate steps to mitigate and effectively manage the risks of harm to individuals as identified by the assessment[277]; and operate a service using proportionate systems and processes to minimise the presence and dissemination of illegal content, the length of time this content is present online, and to swiftly take down illegal content when alerted to its presence[278]. Furthermore, providers must specify in terms of service how individuals will be protected from illegal content and ensure that the terms of service are clear, accessible and consistently applied[279]. Another obligation is that providers must operate reporting systems and complaints procedures so that 'appropriate action' can be taken[280].

Second, particularly with regard to services likely to be accessed by children, providers will additionally have to conduct a children's risk assessment[281], take proportionate steps to mitigate and effectively manage the risks of harm to children, as identified in the assessment, and mitigate the impact of harmful content present on the service[282]. They must also operate a service using proportionate systems and processes to prevent children from encountering harmful content[283]. In their terms of service, it must be specifies how children will be prevented from encountering harmful content[284].

Third, as for the largest online platforms (Category 1)[285], the draft Bill obliges them to conduct an adults' risk assessment and specify in terms of service how harmful priority content to adults and how other harmful content identified through the assessment would be dealt with[286].

[276] Clause 7.

[277] Clause 9(2).

[278] Clause 9(3).

[279] Clauses 9(4) and 9(5).

[280] Clause 15.

[281] Clauses 7(3) and 7(4).

[282] Clause 10(2).

[283] Clause 10(3).

[284] Clause 10(5).

[285] 'Category 1 threshold conditions' relating to a service's number of users and functionalities would be set out in Regulations made by the Secretary of State for Digital, Culture, Media and Sport (see Schedule 4 of the draft Bill).

[286] Clause 11(2). Under Clause 11(3)

Furthermore, Clauses 13 and 14 prescribe that services must protect content of democratic importance as well as journalistic content.

Finally, with regard search engines (search services) their obligations under the draft Bill are similar to those prescribed above in relation to 'user-to-user' services.

Ofcom will be entrusted with enforcement tasks and must prepare codes of practice to assist service providers comply with their duties. Ofcom's powers will include: the issuance of technology notices requiring the use of accredited technology to identify and take down terrorist content and content relating to child sexual exploitation and abuse[287]; information gathering, through information notices, to assist with its online safety functions[288]; the issuance of enforcement notices setting out what a provider or individual must do to comply with the legislation[289]; the issuance of fines up to 18 million GBP or 10 % of qualifying worldwide revenue (whichever is higher) for non-compliance[290]; and business disruption measures[291].

In July 2021, a Joint Committee of both Houses was set up with the task to scrutinise the Bill before a final version is formally introduced to Parliament. In its report from December 2021, the Bill was deemed to be a 'key step forward' in bringing accountability and responsibility to the internet[292]. Despite government efforts to pass the legislation before the House of Commons breaks up for the summer, the government ran out of parliamentary time to push it through and the bill has been shunted to the autumn.

[287] Part 4, Chapter 4.

[288] Part 4, Chapter 5.

[289] Part 4, Chapter 6.

[290] Clause 85. This penalty however must be appropriate and proportionate to the online service provider's failures.

[291] Clause 91.

[292] See https://committees.parliament.uk/committee/534/draft-online-safety-bill-joint-committee/publications/.

SECTION III. REGULATORY OVERSIGHT AND RELATED ACTORS IN THE EU AND THE UK

Online content moderation is increasingly dependent upon a wider number of actors. These involve online platforms and law enforcement authorities cooperating at the national and transnational level, but also regulatory authorities and oversight bodies participating in and contributing to the design, interpretation, monitoring and enforcement of the laws and policies under which content moderation takes place.

The responsibility for actions and decisions related to content moderation— e.g., online content monitoring, collection, processing, and exchange of data, removal of content, account suspension, etc.— **must be distinguished from the responsibility for setting law, policies, and guidelines governing those actions and decisions**. These responsibilities are separate from the responsibility to oversee compliance with online content moderation regulatory duties[293].

The responsibility to create regulations, policies and guidelines, as well as to lay down standards and benchmarks and to develop a robust governance framework remains with the legislators[294]. Part of the legislator's responsibility is also to create **mechanisms for regulatory oversight and ensure accountability**. Such mechanisms take the form of public authorities entrusted with the responsibility to oversee that, in the implementation of online content moderation regulations, and in the exercise of related duties and tasks, online platforms comply with applicable laws and policies, including relevant codes of conducts.

A key characteristic of these regulatory oversight and accountability bodies is that of being independent from both regulators and online platforms providers. This is to ensure accountability of platforms and of being endowed with sufficient investigative and enforcement powers, including the power to impose corrective measures and financial sanctions in case where non-compliance from regulatory duties is detected. **Regulatory oversight by public authorities can also be complemented by independent civil society organisations**[295].

Given that the implementation of online content moderation activities entails compliance with **a multi-level normative and policy framework** — encompassing data protection, provision of services in the digital marketplace, media laws, copyright protection, consumer protection, competition law, etc.— **a mosaic of bodies currently comprise the framework of regulatory oversight that apply to online content moderation**.

---

[293] Bayer, J. et al. (2021), The fight against disinformation and the right to freedom of expression, Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies PE 695.445, July 2021, p. 62.

[294] Kaye, D. (2019), 'A New Constitution for Content Moderation', *OneZero*.

[295] Ibid.

These oversight structures complement other accountability mechanisms and venues that apply to online content moderation, which also include **judicial redress mechanisms**. These remain competent to ultimately review platforms' decisions implementing online content moderation (pursuant to regulations and/or based on their own Terms of service), but also regulatory authorities' decisions related to online platforms' regulatory compliance.

III.1. OVERSEEING THE IMPLEMENTATION OF NORMS AND POLICIES ON ILLEGAL AND HARMFUL CONTENT ONLINE IN THE EU

III.1.1. Regulatory oversight by EU data protection authorities: Roles and limitations

National data protection authorities (DPAs) are responsible to supervise through investigative and corrective powers the application of EU data protection law, to provide expert advice on data protection issues, and to handle complaints lodged against violations of the General Data Protection Regulation (GDPR) and relevant national laws.

Under the GDPR, the principle of accountability[296] entails that **the data holder (controller or processor) is accountable for ensuring compliance with the regulation's principles, provisions, and associated data subjects' rights**. This principle also applies to data processing activities undertaken by private platforms (acting as controllers or processors) in the implementation of content moderation duties pursuant to national or EU law.

While not all forms of content moderation necessarily require processing of personal data, they often do. In particular, they entail processing of user-generated content by content moderators, as well as the design and testing of content moderation algorithms, and the implementation of the automated decision-making tools which, increasingly, are deployed by online platforms to monitor, detect, flag, or black or remove, illegal or harmful content. As **the DPAs accountability framework** applies to automated decision making in the scope of the GDPR[297], it **also extends to the use of AI in the online content moderation processes.**

The widening scope and extent of online platforms' content moderation activities, including through automated means, has indeed led DPAs to **growing concern about online platforms' compliance with data protection rights and safeguards**. Concerns related to online content moderation taking place in a wider context of 'endemic monitoring' of individuals' behaviour by private platforms recently led the European Data Protection Supervisor (EDPS) to stress that **content moderation should take place in accordance with the rule of law, and that content moderation measures should be as targeted as possible, in accordance with principles of**

[296] Art. 5(2) of the GDPR.

[297] Castets-Renard, C. (2019), 'Accountability of Algorithms in the GDPR and beyond: A European Legal Framework on Automated Decision-Making' (2019) 30(1) Fordham Intellectual Property, Media & Entertainment Law Journal 91, at 107.

necessity, proportionality, and data minimisation[298]. The EDPS also stated that content moderation should, by design and by default, not involve the processing, collection and disclosure of personal data[299].

DPAs across the EU are increasingly dealing with issues related to content moderation by online platforms. In Italy, for instance a special division of the national DPA was created to deal exclusively with content moderation issues[300]. A particularly effective means to ensure consistency of application of the GDPR is provided by DSAs' 'corrective power' to 'impose a temporary or definitive restriction including a ban on processing'[301]. The GDPR also expressly requires that national DPAs 'cooperate with each other and, where relevant, with the Commission' through a 'consistency mechanism' designed to ensure consistency of application of data protection laws throughout the EU[302]. These enforcement instruments and coordination mechanisms can be used to monitor and enforce compliance with data processing law, and related guarantees, in the content moderation context.

Relevant regulatory oversight activities performed by the DPAs also involve the provision of guidance, in particular with regard to the protection of data subjects rights vis-à-vis automated data processing, which is also relevant in the context of content moderation[303]. In this specific regard, the EDPS has underlined that profiling for purposes of content moderation should be prohibited unless the provider can demonstrate that such measures are strictly necessary to address forms of systemic risks, as explicitly identified by the DSA[304].

Another DPAs duty that is the follow up of user's complaints against content moderation data processing activities[305]. Under the GDPR, individuals confronted with a data protection infringement— including those arising from online content moderation activities— have the right to turn directly to the judiciary[306]. At the same time, they also have the right to lodge a complaint with a national DPA[307]. DPAs are obliged to facilitate the submission of complaints[308]. They are tasked with the duty to handle lodged complaints and with investigating, to extent appropriate, the complaints' subject matter. Data subjects have the right to an effective

[298] EDPS, Opinion 1/2021 on the Proposal for a Digital Services Act, para. 25.

[299] Ibid. paras 25 and 52.

[300] Global Forum of Freedom of Expression (2021), "Italian Data Protection Authority v. TikTok", Columbia University, Global Forum of Freedom of Expression.

[301] Art. 58(2)(f) of the GDPR.

[302] Art 63 of the GDPR.

[303] Article 29 Working Party (2017), Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679. During its first plenary meeting the European Data Protection Board endorsed the Article 29 Working Party Guidelines.

[304] EDPS, Opinion 1/2021 on the Proposal for a Digital Services Act, para. 26.

[305] González Fuster, G. et al. (2022), 'The right to lodge a data protection complaint: OK, but then what? An empirical study of current practices under the GDPR', Data Protection Law Scholars Network (DPLS), June 2022.

[306] Art. 79 of the GDPR.

[307] Art. 77 of the GDPR.

[308] Art . 57(2) of the GDPR.

remedy against the decisions of DPAs, as well as in case of lack of action or lack of information about the progress or outcome of their complaints[309]. Complaints might also be handled by multiple DPAs, cooperating through the one-stop-shop mechanism[310].

National DPAs provide a crucial network of regulatory oversight and external accountability framework covering data processing (regardless of whether it is voluntary, or legally mandated) in online content moderation. Importantly, DPAs are responsible for monitoring the implementation of online content surveillance and moderations tasks by national law enforcement authorities, and to ensure the latter's compliance with the data protection framework provided under the Law Enforcement Directive[311].

First, it must be clarified that it is not necessary that the same authority will be assigned as DPA for the purposes of the GDPR and for the purposes of the LED. In practice, in most Member States the same authority is assigned to supervise both the GDPR and the national laws transposing the LED. But there exists lack of consistency between the provisions of the LED and those of the GDPR on oversight. In particular, the powers conferred to DPAs under Article 47 of the Directive are not aligned with those listed in the GDPR and are more limited without justification[312]. For example, the power to impose 'administrative fines' or penalties or to suspend data flows to a recipient in a third country, or to an international organisation are absent[313]. Investigatory powers are also more limited[314].

DPAs' regulatory oversight and external accountability role complements, and is complemented by, judicial oversight mechanisms and related remedies. This kind of networked and multilevel regulatory oversight and external accountability infrastructure, which is designed to guarantee effective implementation of data protection law, does not exist in other EU policy areas or legal domains under which content moderation takes place. However, Task Force discussions confirmed that there also are clear limitations to the DPAs capacity to effectively oversee online platforms activities, and to prevent or redress the different abuses potentially arising from implementation of online content moderation as regulated under different pieces of EU and national legislation.

---

[309] Art. 78 of the GDPR.

[310] Art. 60 of the GDPR.

[311] Directive (EU) 2016/680 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 by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L 119/ 89 (Law Enforcement Directive – LED).

[312] See Caruana, M., (2017), 'The reform of the EU data protection framework in the context of the police and criminal justice sector: harmonisation, scope, oversight and enforcement', *International Review of Law, Computers and Technology*, Vol. 33, No 3, pp. 249, 259.

[313] Compare Article 47 of the Directive with Article 58 of the GDPR.

[314] In particular, the power to obtain from the controller, or the processor, access to ' any premises of the controller and the processor, including to any data processing equipment and means' (GDPR Art.58(1)(f)) is not listed in the Directive.

In the first place, an intrinsic limitation derives from **the very nature and scope of the DPAs mandate**, which exclusively relates to **the effective and consistent application of EU or national laws in the field of data protection**. Such a duty is also essential in the perspective of data protection remedies, which are directly linked to the right to the protection of personal data, and the right to an effective judicial remedy, enshrined respectively in Article 8 and 47 of the EU Charter of Fundamental Rights. While it is useful to keep under check private and public data processing activities that can be detrimental to data protection**, the DPAs focus on data protection is not sufficient to prevent the occurrence of other legal challenges** potentially arising from online content moderation, including for instance with regard to **freedom of expression and association, non-discrimination**, etc. DPAs do not have **the competence and perhaps even the expertise to consider the effective respect of such rights and freedoms**.

Another critical limitation derives from **the lack of adequate resources or capacity to ensure effective monitoring of correct implementation of the GDPR and/or the LED**. In some EU countries, DPAs have been the target of criticisms in this regard. The Irish Data Protection Commission has been especially criticised for its delay in taking enforcement action under the GDPR against large online platforms with European headquarters in Ireland (e.g., Apple, Google, Facebook, Twitter). Such criticisms were expressed, for instance, by Germany's Federal Commissioner for Data Protection[315], the European Commission's Vice-president Vera Jourova[316], and CJEU's Advocate General Bobek[317]. In May 2021, the European Parliament passed a resolution calling on the Commission to launch an infringement procedure against Ireland for failing to enforce the GDPR[318]. In November 2021, the Irish Council for Civil Liberties filed a complaint with the European Ombudsman over the European Commission's failure to initiate infringement procedures against Ireland over its application of the GDPR[319]. At the same time, the Irish DPA is clearly not the only authority that remains weak, including most notably in terms of resources and expertise available.

DPAs' capacity to effectively oversee cross-border data processing[320] in the context of online moderation can also be affected by **jurisdictional factors**. Issues arise when national DPAs have to deal with online platforms' data processing activities involving **data transfers to law enforcement authorities from non-EU countries, such as the United States, upon which the Commission has not adopted an adequacy decision**. At the transatlantic level, the EU and the

[315] Scally, D. (2020), 'German regulator says Irish data protection commission is being "overwhelmed"' , *The Irish Times*, 3 February 2020.

[316] Weckler, A. (2021), 'EC vice president takes aim at "understaffed" Irish Data Protection Commission: Vera Jourova said a "harmonised approach" should be pursued around the EU', *Independent.ie*, 2 November 2021.

[317] Opinion Of Advocate General Bobek delivered on 13 January 2021(1), Case C-645/19, Facebook Ireland, Limited, Facebook Inc., Facebook Belgium BVBA.

[318] Bertuzzi, L. (2021), 'MEPs call for infringement procedure against Ireland', *Euractive*, 20 May 2021.

[319] Ó Fathaigh, R. (2021), 'The Digital Services Act proposal and Ireland', *DSA Observatory*, 3 December 2021.

[320] Art. 4(23) of the GDPR.

United States have committed to a Trans-Atlantic Data Privacy Framework aimed at establishing a new legal mechanism for transfers of EU personal data to the US.

The need to establish such mechanisms arose after the CJEU decision in the *Schrems 2* case[321], which struck down the Commission's adequacy decision underlying the EU-US Privacy Shield framework. The judgment was motivated by **the lack of safeguards, on the US side, ensuring that intelligence activities are necessary and proportionate** in the pursuit of national security objectives, and by **the absence of a mechanism for EU individuals to seek effective redress** if they believe they are unlawfully targeted by US intelligence authorities data processing activities.

Finally, critical limitations derive from the fact that **pieces of EU legislation such as TERREG** which regulates various forms of private-public cooperation in the online content moderation domain, **does not expressly foresee a role for DPAs**. In particular, TERREG refers to one or more 'competent authorities' entrusted with tasks that relate to issuing and supervision of orders for content removal, without judicial supervision and enforcement of the Regulation. Other rules relating to supervision do not exist. In fact, Article 12 and Recital 35 state that the fact that Member States must ensure that their competent authorities have the necessary powers and sufficient resources to fulfil their obligations under this Regulation should not prevent supervision **in accordance with *national constitutional law***. As a result, **supervision in this context must take place in line with national law**, which will necessarily create **lack of consistency in the approaches of different Member States** and differentiated standards depending on resources and expertise.

In turn, the above-mentioned recent Commission proposal for a Regulation to prevent and combat child sexual abuse has taken a different approach. Chapter III of the proposal is dedicated to supervision requiring Member States to designate one or more competent authorities as responsible for the application and enforcement of the relevant rules[322]. One of these authorities will operate as Coordinating Authority for child sexual abuse issues[323] and must operate in an objective, impartial, transparent, and timely manner[324]. Supervision of that Coordinating Authority is not precluded and must take place under national law to the extent that such supervision does not affect their independence[325].

[321] Court of Justice of the European Union Judgment in Case C-311/18 Data Protection Commissioner v Facebook Ireland and Maximillian Schrems, of 16 July 2020.

[322] Article 25(1).

[323] Article 25(2).

[324] Article 26.

[325] Article 26(3).

III.1.2 Regulatory authorities as enforcers of online content moderation

The legal and institutional context for the regulatory oversight of online content moderation is complex and varied among Member States. In parallel with normative and policy developments, **national administrative authorities have been tasked with new regulatory oversight duties**, as well as with **powers to directly enforce content moderation laws and policies** vis-à-vis online platforms.

In France, the *Conseil Supérieur de l'Audiovisuel* (i.e., the media regulator) expressly stated that, under the Law on Combatting the Manipulation of Information[326], it reserves itself with the 'ability [to] request […] any information should it observe a manipulation of information or an attempt to manipulate information likely to disturb public order or to affect the sincerity of an election'[327]. In 2018, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom and Expression raised express concerns about the cooperation mechanisms established under this French Law, in as far as it authorised government authorities' access to subscribers' data related to reported content, without prior court approval. In particular, it was noted that the 'absence of a judicial authorisation for the disclosure of personal information would be a restriction which is neither targeted nor protective of rights to a fair hearing and would therefore not meet the strictest for imposing restrictions on privacy and freedom of expression'[328].

Similar initiatives proliferated across different EU countries. In Romania, a Presidential Decree permitted the communications regulator to directly order the removal of and block access to online content that 'promotes false news' regarding Covid-19 protection and prevention measures[329]. In the Netherlands, a draft law[330] envisages the creation of a new administrative authority (the ZBO) that would be responsible to oversee a general obligation (or duty of care) for hosting providers to take 'suitable and proportionate measures' to limit the storage and distribution of child sexual exploitation material through their services. The authority would also play an active role in data sharing and the management of 'hash sharing collaborations'. The proposal also envisages entrusting the new authority with the power to directly order removal of child sexual abuse material hosted by communication services established in the Netherlands. Infringements of these obligations are subject to potential administrative penalties, including fines, which would also be imposed by the authority[331].

---

[326] Loi no 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation d' l'information.

[327] Conseil Supérieur De L'audio Conseil Supérieur De L'audiovisuel visuel (2019), para 4(a).

[328] U.N. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion, and Expression, U.N. Doc. OL/FRA 5/2018, 28 May 2018, p. 7.

[329] Decret semnat de Președintele României, domnul Klaus Iohannis, privind instituirea stării de urgență pe teritoriul României, 16 martie 2020.

[330] 'Wetsvoorstel bestuursrechtelijke aanpak online kinderpornografisch materiaal'.

[331] Van Hoboken, J. (2021), 'The DSA proposal and the Netherlands', DSA Observatory Analysis, 21 October 2021.

In Ireland, the Online Safety Media Regulation Bill[332] aims at creating a new regulatory framework under which an Online Safety Commissioner would be established as part of a new Media Commission. The Commissioner would be empowered to issue 'binding online safety codes', but also have regulatory compliance, enforcement, and sanction powers. These would include the power to 'remove or restore individual pieces of content'. Critically, it has been noted how the Bill would not only apply to 'material which is a criminal offence to disseminate under Irish [or Union law]', but also other 'harmful content', such as material 'which is likely to have the effect of intimidating, threatening, humiliating or persecuting a person to which it pertains'[333]. The government has also made recommendations for disinformation to also be included as a category of harmful online content. Civil society actors argued that the new law would create 'vague new online-only offences', and may not satisfy the 'standards of legality, necessity and proportionality'. It has also been noted that, by allowing the issuing of notices for removal of content with a 'threshold so low that it could seriously damage individuals', the law could also endanger the effective protection of the constitutional right to freedom of expression[334].

A critical common feature of these proposals or initiatives is that **they empower administrative authorities with the competence to directly issue 'takedowns orders', with no guarantee or requirement for *ex ante* judicial review on or approval of removal orders**, which raises challenges from the perspective of effective protection of the right to freedom of expression.

The creation of these authorities at the national level, respond to some extent to the need to ensure compliance with EU law obligations to fight against illegal content, for instance as required under the TERREG regulation. As mentioned earlier, the latter requires national competent authorities to *inter alia* issue removal orders without any requirement for intervention and oversight by a judicial authority. At the same time, national initiatives empowering administrative authorities with the competence to impose fines and directly issue removal orders (in some case, without need for judicial supervision) could generate **incoherencies with the upcoming DSA**. Contrary to the TERREG, **the DSA refers to orders to which providers must respond without undue delay** that are issued by either administrative or judicial authorities in accordance with EU or national law and in compliance with EU law.

[332] Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media (2020), Online Safety and Media Regulation Bill, January 2020.

[333] Ó Fathaigh, R. (201), 'The Digital Services Act proposal and Ireland', *DSA Observatory*, Analysis, 3 December2021.

[334] Cronin, O. (2021), ICCL submission on the Online Safety and Media Regulation Bill, 8 March 2021.

III.2. REGULATORY OVERSIGHT UNDER THE DSA

The DSA, as discussed in *Section II.3.2*. above, aims to ensure unity of its enforcement through strong cooperation mechanisms between the national competent authorities and the Board and the Commission[335]. As explained in *Section II* of this Report, the DSA appoints the Digital Services Coordinator of the Member State of establishment as the main supervision and enforcement authority alongside other competent authorities that may be designated at national level[336].

The Digital Services Coordinator (DSC) function did not exist before, so at the national level it is still unclear who will be appointed as the DSC. Indeed, at the national level, **the legal and institutional context of platform regulation is complex and varied**, involving different administrative bodies that may be designated as DSCs. These may be new independent administrative authorities created under new laws (see, for example the Netherlands and Ireland), or existing supervisory authorities (see, for example, Germany).

Beyond the body that will be assigned as the DSC (and additional competent authorities), **their hierarchical relationship with other independent regulatory bodies,** and at which instances competent authorities would be performing tasks related to the DSA and when related to other laws, are issues that **have not been addressed**. Article 38(2) of the DSA enables Member States to maintain the assignation of specific tasks or sectors to other authorities and allows Member States discretion to provide for cooperation mechanisms and regular exchanges of views of the Digital Services Coordinator with other national authorities where relevant for the performance of their respective tasks.

Member States will also be able to designate other competent authorities for the enforcement of the DSA, for example for specific sectors, such as electronic communications' regulators, media regulators or consumer protection authorities (Recital 72 and Article 38). Whereas the DSA sets out the powers of the Coordinators, **it pays little explicit attention to other relevant enforcement authorities to address situations of potential overlap in competencies and lacks institutionalised and structured cooperation** between other competent oversight authorities in matters of mutual concern. One common criticism in this regard is **that the DSA fails to provide a clear legal basis for the DSCs, the EBDS, and the Commission to cooperate and/or consult with other EU enforcement networks**[337].

How do the enforcement mechanisms explained above interplay with existing enforcement structures established under other acts of the EU? **The involvement of other regulatory bodies is generally discretionary**: **the DSA leaves it at the Member State discretion to provide for**

[335] Articles 38, 44, 44a, 44b, 45, 46 of the DSA.

[336] Articles 38 and 40 of the DSA.

[337] Zeybek, B., and van Hoboken, J. (2022), 'The Enforcement Aspects of the DSA, and its Relation to Existing Regulatory Oversight in the EU', 4 February 4, p. 202, https://dsa-observatory.eu/2022/02/04/the-enforcement-aspects-of-the-dsa-and-its-relation-to-existing-regulatory-oversight-in-the-eu/.

regular exchanges of views with other authorities. The EBDS may invite other national authorities to meetings (Article 48(1) of the Commission) and may cooperate with other EU bodies, offices, agencies, and advisory groups (as well as external experts as appropriate (Recital 91 and Article 48(5) of the Commission). The Task Force discussions concluded that **the already established multiplicity of actors and their interplay may create confusion and require future streamlining on the basis of operational challenges arising in practice**.

III.3. OVERSEEING THE IMPLEMENTATION OF NORMS AND POLICIES ON ILLEGAL AND HARMFUL ONLINE CONTENT IN THE UK

As Smartt (2014) notes, in the UK there are **three regulatory models**: industry self-regulation, co-regulation (a combination of industry self-regulation and with oversight by a statutory body), and statutory regulation control by a statutory authority, such as Ofcom[338]. This sub-section provides an overview of the relevant regulators, first discussing the Information Commissioner's Officer (ICO) (*Section III.3.1*), before focussing in more detail on the Ofcom, which will be the regulator according to the Online Safety Bill (*Section III.3.2.)*, and the Digital Regulation Cooperation Forum (*Section III.3.3.*).

III.3.1. The Information Commissioner's Office

The Information Commissioner's Office (ICO) was set up to uphold information rights in the public interest, promoting openness by public bodies and data privacy for individuals. It is an executive non-departmental public body, sponsored by the Department for Digital, Culture, Media & Sport. As a regulator, the ICO covers a growing amount of legislation most notably the Data Protection Act 2018, the Freedom of Information Act 2000, and the Privacy and Electronic Communications Regulations 2003.

Indeed, The ICO is responsible for enforcing data protection legislation and monitoring organisational compliance, and to those ends has the power to impose fines for non-compliance. It represents individuals and has a proactive role in raising awareness and informing the public about privacy-related aspects so as to make informed decisions about how their personal data is used. Although ICO operates independently in the exercise of their statutory functions, some issues require the approval of the Secretary of State such as funding and the level of fees charged to data controllers.

In view of its mandate, with regard to online content, the Information Commissioner has issued an Age Appropriate Design Code setting out standards that online services need to follow around children's personal data[339]. The code is ground-breaking; it focuses on **a 'by design approach' in order to protect children online**. All major social media and online services used by children in the UK will need to conform to the code, giving the impact an international reach.

---

[338] Smartt, U. (2014), *Media and Entertainment Law*, Routledge, p. 528.

[339] Information Commissioner's Office (2021), 'Introduction to the Age appropriate design code', https://ico.org.uk/for-organisations/guide-to-data-protection/ico-codes-of-practice/age-appropriate-design-code/.

Furthermore, as will be discussed below, **the authority cooperates with other regulators to bring forward and ensure a coordinated and consistent approach on various issues of common interest**.

III.3.2. Office of Communications

The Office of Communications (Ofcom) is the regulator and competition authority for the UK communications industries. It regulates the TV and radio sectors, fixed line telecoms, mobiles, postal services, plus the airwaves over which wireless devices operate. It was created in 2003 and operates under statute, the Communications Act 2003, which outlines the regulator's general responsibilities, further enhanced by the Digital Economy Act 2010.

Section 3 of the Communications Act 2003 describes Ofcom's duties as to further the interests of citizens in connection to communications matters and to further the interests of consumers in relevant markets, where appropriate by promoting competition. Its responsibilities fall within six main areas, which include the application of adequate protection for audiences against offensive and harmful material or against unfairness or the infringement of privacy[340].

Ofcom is also responsible for the specification of the procedural and enforcement aspects of obligations to providers through the approval or adoption of legally binding codes of practice. Therefore, pending the adoption of the Online Safety Bill, the authority does not handle content written or posted in online platforms. However, it does have responsibilities in relation to video sharing platforms, such as TikTok and Snapchat: in that regard, it has issued Guidelines for platforms on measures to protect users of harmful content[341]. A Plan and Approach to the VSP rules has also been published which outlined five initial regulatory priorities: reducing the risk of child sexual abuse material on adult sites; laying the foundations for age verification on those sites; tackling online hate and terror; ensuring an age appropriate experience on platforms popular with under-18s; and ensuring VSPs' processes for reporting harmful content are effective[342]. Ofcom works with the Department for Digital, Culture, Media & Sport, is independent and is funded by fees paid by the companies regulated.

The passing of the Online Safety Bill will bring service providers within the regulatory and enforcement powers of Ofcom. From a fundamental rights perspective, **Ofcom provides a model for regulatory agency, which is targeted at communications and has a wider mandate that includes but goes beyond the rights of privacy.** However, it will have **a difficult balancing**

[340] Other areas are: Ofcom ensures that individuals are able to use communications services, including broadband; that a range of companies provide quality television and radio programmes that appeal to diverse audiences; the universal postal service covers all UK addresses six days a week, with standard pricing; and the radio spectrum is used in the most effective way.

[341] https://www.ofcom.org.uk/__data/assets/pdf_file/0015/226302/vsp-harms-guidance.pdf.

[342] Ofcom (2021), 'Video Sharing Platforms: Ofcom's Plan and Approach', https://www.ofcom.org.uk/__data/assets/pdf_file/0016/226303/vsp-plan-approach.pdf.

act to manage the different interests at stake, and to be an effective but not overly heavy-handed regulator[343].

In preparation for the appointment, since December 2020, Ofcom has been funded by Government to strengthen its capabilities[344]. In July 2022, it issued a Roadmap for implementation of its duties under the Bill which are summarised as falling into four streams: protecting people from illegal content; protecting children in particular from age inappropriate content; empowering adults to protect themselves from legal but harmful content; and increasing public transparency of services' actions to protect users[345]. The publication of the Roadmap may be seen as an effort on behalf of Ofcom to give the earliest possible warning to businesses as to when they will be asked to engage with consultations[346].

As a priority action, Ofcom will focus on child sexual abuse material and grooming, terrorism and online fraud, and limiting harmful content to children[347]. Regulation of (the particularly controversial) 'legal but harmful' content is to be phased in at a later time following secondary legislation. Furthermore, Ofcom intends to 'prioritise engagement with high-risk or high-impact services to understand their existing safety systems and how they plan to improve them', adopting a 'risk-based 'supervisory' approach'[348]. Furthermore, its engagement will go beyond the higher volume and higher services which are covered by the scope of the Online Safety Bill to cover 'some smaller services […] posing particular risks of harm by virtue of their content offer, userbase or service features'[349].

That said, the powers of Ofcom in the Online Safety Bill have been criticised from the perspective that the Bill gives broad power to the Secretary of State over its implementation and can endanger Ofcom's independence. In particular, section 53 provides that the Secretary of State will define priority content considered harmful to children and adults in secondary legislation. In that regard, concerns have been raised that, in practice, putting secondary legislation before Parliament will merely be a rubber-stamping exercise without the scrutiny required[350].

---

[343] 'Online Safety Bill: first indications of Ofcom's regulatory approach', Panopticon blog, 6 July 2022, https://panopticonblog.com/2022/07/06/online-safety-bill-first-indications-of-ofcoms-regulatory-approach/.

[344] Ofcom (2022), 'Online Safety Bill: Ofcom's roadmap to regulation', https://www.ofcom.org.uk/__data/assets/pdf_file/0016/240442/online-safety-roadmap.pdf.

[345] Ibid.

[346] Panopticon blog (2022), 'Online Safety Bill: first indications of Ofcom's regulatory approach', 6 July, https://panopticonblog.com/2022/07/06/online-safety-bill-first-indications-of-ofcoms-regulatory-approach/

[347] Government has recently proposed amendments to the Online Safety Bill which would further strengthen Ofcom's powers relating to child sexual abuse material, including an ability to require providers to use or develop technology to identify such content in user communications.

[348] Ofcom (2022), Online Safety Bill: Ofcom's roadmap to regulation, op. cit. 8.

[349] Ibid.

[350] Article 19.

Another point of concern is that in accordance with sections 39 and 40, Ofcom will be required to submit its code of practice to the Secretary of State which can direct Ofcom to modify the draft on vaguely-defined grounds such as 'for reasons of public policy'. Furthermore, section 78 enables the Secretary of State to set out a statement of the Government's strategic priorities, which Ofcom will have to consider when carrying out its functions. Perhaps then the Roadmap may be seen as an effort from the authority to pre-empt prioritisation. Finally, according to section 147, the Secretary of State may provide guidance about how Ofcom should exercise its powers.

Notwithstanding the fact that the Joint Committee on the Draft Online Safety Bill recommended constraining its powers over the Bill, the aforementioned instances demonstrate how **Ofcom, which is an independent regulator, will be overly dependent upon the Secretary of State**, which undermines its status and will affect its **regulatory oversight and overall performance of its duties**. As the Chair of the Digital, Culture, Media and Sport Committee stated, 'A free media depends on ensuring the regulator is free from the threat of day-to-day interference from the executive. The Government will still have an important role in setting the direction of travel, but Ofcom must not be constantly peering over its shoulder answering to the whims of a backseat-driving Secretary of State'[351].

To date, the Online Safety Bill has added provisions for Ofcom on other matters: the authority will be able to recommend the use of tools for content moderation, user profiling and behaviour identification with strong safeguards to ensure these are used only where necessary and where it is proportionate to the harms posed[352]. Additional amendments—essentially leaving out clauses 39 and 40 — have been made by the Digital, Culture, Media and Sport Committee with the aim to limit the powers of the Secretary of State to interfere in the work Ofcom[353].

III.3.3. The Digital Regulation Cooperation Forum

To achieve maximum effect, Ofcom cooperates with other relevant regulators. A paradigmatic example in this context is the creation of **the Digital Regulation Cooperation Forum (DRCF)**, which brings together **the major UK regulators tasked with regulating digital services**: (a) the Competition and Markets Authority (CMA); (b) the Financial Conduct Authority (FCA) (c) the Information Commissioners Office (ICO), and (d) Ofcom, each regulating in their respective

[351] 'MPs propose amendments to Online Safety Bill to ensure Ofcom independence', 4 July 2022, https://committees.parliament.uk/committee/378/digital-culture-media-and-sport-committee/news/171833/mps-propose-amendments-to-online-safety-bill-to-ensure-ofcom-independence/.

[352] See https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-factsheet.

[353] House of Commons (2022), 'Digital, Culture, Media and Sport Committee, Amending the Online Safety Bill (First Report of Session 2022–23)', 30 June, https://committees.parliament.uk/committee/378/digital-culture-media-and-sport-committee/news/171833/mps-propose-amendments-to-online-safety-bill-to-ensure-ofcom-independence/.

areas of expertise. This formal construct builds on **strong operational relationships among the regulators with a distinct focus on the digital landscape.**

Within this framework, in 2021 to 2022 the Information Commissioner and Ofcom strengthened their existing cooperation on online safety, following the introduction of the ICO's Age Appropriate Design Code (AADC) and Ofcom's regulation of UK-established Video-Sharing Platforms (VSPs)[354]. In fact, the two regulators worked together to develop respective guidance, such as Ofcom's guidance to VSPs and the ICO's Opinion on the use of age assurance technologies. The have also commissioned joint research exploring children and parents views on age assurance online.

The operation of the DRCF may provide some insights into how the European Board for Digital Services under the DSA may cooperate in practice. Working across the different related spheres of an increasing digital environment and identifying synergies is a highly effective way of collaboration to ensure coherence in approach and different applicable rules, and is testament to emphasis on the need for a greater level of coordination and cooperation.

Task Force discussions confirmed that **prioritising coordination in different lines of inquiry between different regulators ensures maximised effectiveness, influence, and impact**. Moreover, it leads to a **more efficient development and retaining of the right skills, knowledge, expertise, and organisational capability** to deliver effective digital regulation for both individuals and businesses. Identifying and clarifying the linkages in different areas of work to tackle common challenges through a formalised structure also **enables the sharing of data that needs to be shared**.

Despite the undeniable benefits of this approach **there are still certain challenges that call for caution. This is particularly so when considering the potential transplantation of this model at the regional and transnational level**. On the one hand, the prioritisation of the different tasks and areas of interest across the different regulators may be challenging (though for now there appears to be some consensus about the importance of child sexual abuse related content and counterterrorism, which is in line with the current EU approach). On the other hand, some of the presentations and discussions held during Task Force meetings revealed that a much more difficult endeavour is **to ensure coordination among different supervisory authorities at the transnational level,** taking into account the different degrees of digitalisation of services, resources, and potentially different areas of concern.

354 https://www.gov.uk/government/publications/digital-regulation-cooperation-forum-annual-report-2021-to-2022/digital-regulation-cooperation-forum-annual-report-2021-to-2022.

SECTION IV. FUNDAMENTAL RIGHTS AND RULE OF LAW CHALLENGES

Increasing reliance on private platforms to moderate online content through various forms of pre-moderation (exercised before content is posted) and post-moderation (exercised after content publication) has serious **fundamental rights implications and impacts**. Impact on different fundamental rights depends on the specific ways in which content moderation takes place, as well as on the articulation of content moderation duties and their implementation through various forms of private-public cooperation. This Section of the Report examines the challenges emerging from online content moderation in relation to privacy and data protection (*Section IV.1.*), freedom of expression (*Section IV.2.*) and the rule of law, due process, and effective remedies (*Section IV.3.*).

IV.1. PRIVACY AND DATA PROTECTION

The monitoring and surveillance of the online environment for the purpose of content moderation (including through access to, and processing of, large amounts of various categories of users' personal data by a wide range of private and public actors) raises concerns in relation to the protection of the rights to privacy and data protection, as safeguarded under the EU legal system and international and regional human rights law. Privacy and data protection are inherently embedded in the right to human dignity[355]. They are however distinct fundamental rights under the Charter, enshrined in Article 7 and 8 respectively. Privacy is broader in the sense that it encompasses a multiplicity of aspects apart from the processing of personal data, whereas data protection falls largely within the aspect of privacy.

IV.1.1. Initiatives enabling forms of generalised monitoring of online content

Of particular concern for the rights to privacy and data protection are **those national or supranational initiatives that require companies to track or trace content across their platforms**, prospectively and retrospectively, and **across jurisdictional boundaries**.

Article 15 of the e-Commerce Directive presents an express prohibition on Member States imposing on providers a general obligation to 'monitor the information they transmit or store', or to 'actively seek facts or circumstances indicating illegal activity' in the online space[356]. Under the Directive, Member States are not precluded from imposing injunctions for providers to inform competent public authorities of alleged illegal activities undertaken[357], but these should

[355] Van Hoboken, J., Ó Fathaigh, R. (2021), 'Regulating Disinformation in Europe: Implications for Speech and Privacy', *UC Irvine Journal of International, Transnational and Comparative Law*, Vol. 6; Ibid, (2021) 'Symposium: The Transnational Legal Ordering of Privacy and Speech', p. 18.

[356] ECD, Article 15 a).

[357] ECD, Article 15 b).

only be permitted **if they are *specific* rather than *general* and directed at preventing particular infringements in specific cases**[358].

Furthermore, the e-Commerce Directive has allowed Member States to impose 'duties of care' which can reasonably be expected from internet service providers, and which can be specified by national law in order 'to detect and prevent certain types of illegal activities' in the online environment[359]. **The actual scope and content of such duties of care have therefore not been defined precisely**, and it is unclear what differentiates them from an actual generalised monitoring obligation[360]. Such unclarity, coupled with the encouragement for Member States to conclude 'voluntary agreement' (e.g., in the form of codes of conduct) with internet service providers, has consequently led to the proliferation of national initiatives enabling various forms of online content surveillance that service providers have been asked or mandated to perform across their platforms.

The interpretation of the prohibition on general monitoring contained in the e-Commerce Directive has not been universally agreed by scholars. Some authors have argued that 'rightly understood, a prohibited general monitoring obligation arises whenever content— no matter how specifically it is defined— must be identified among the totality of the content on a platform'[361]. This would mean that 'a content moderation duty can only be deemed permissible if it is specific in respect of both the protected subject matter and potential infringers'[362]. An alternative view is that **the generality of the monitoring is not determined by what is being monitored, but by the objective of the monitoring**. This would mean that an obligation to monitor all the information on a platform would be permissible as long as there is a specific type of illegal content identified *ex ante*[363].

Initiatives promoting forms of generalised monitoring have flourished at both the national and supranational level. In the UK, the Online Harms White Paper foresees a duty for platforms to 'prevent known terrorist or Child Sexual Exploitation and Abuse (CSEA) content being made available to users'. While the White Paper makes clear that the duty will apply to private services (including email and messaging services, as well as private storage services), it does not explain how this could be done without proactive monitoring by companies of all content on their platforms, or without requiring companies to directly access content of private communications.

---

[358] ECD, Recital 47.

[359] ECD, Recital 48.

[360] Kuczerawy, A.(2019), 'To Monitor or Not to Monitor? The Uncertain Future of Article 15 of the E-Commerce Directive', KU Leuven CiTiP.

[361] Senftleben, M., Angelopoulos, C. (2020), 'The Odyssey of the Prohibition on General Monitoring obligations on the Way to the Digital Services Act: Between Article 15 of the E-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market', Amsterdam/Cambridge, p. 2.

[362] Ibid., p. 2.

[363] Ibid at p. 8.

The Online Safety Bill raises concerns about the possibility of general monitoring. In essence, in order for online platforms to comply with their duties to prevent individuals from being exposed to certain illegal or harmful content, they will have to monitor all content on their platforms. The Bill allows Ofcom to impose a 'proactive technology' requirement to identify and remove all kinds of illegal content or content that is harmful to children[364]. The general monitoring obligation is even more challenging given that pursuant to Section 2 private channels such as WhatsApp, Signal, or Telegram are also covered under the scope of obligations. Furthermore, Ofcom will have the power to order a provider to use 'accredited technology' to identify child sexual exploitation and abuse content—whether such content is communicated publicly or privately[365]. No distinction is made between public platforms as opposed to private messaging services, which may signify that the latter (which operate on the basis of end-to-end encryption) may be found to violate the Bill[366].

EU Member States have also introduced measures that impose or enable large-scale online surveillance by private platforms. In Germany, the aforementioned NetzDG creates an obligation to assess the legality of content by reference to the German Criminal Code, and to delete any allegedly illegal speech within strict time frames. It has been observed how meeting the deadlines for deletion can hardly be achieved without resorting to some form of general monitoring system[367]. In France, the Law on Combatting the Manipulation of Information[368] encourages platforms to set up 'appropriate procedures allowing for the detection of accounts disseminating false information on a massive scale'.

Generalised monitoring of online content has also been mandated in the context of transnational initiatives (e.g., the Global Internet Forum to Counter Terrorism), or supranational regulations (e.g., CSAM) requiring the filtering of content that—without independent judicial or administrative oversight—has been previously labelled as illegal or harmful. As mentioned above in the case of the Online Safety Bill, **by involving the development by companies of capabilities for proactively screening content, these initiatives could also limit platforms' ability to encrypt private messages,** which disrupts their business model and removes their competitive edge in their respective market.

In the case of the TERREG, there is no rule preventing service providers from taking up general monitoring activities and there is no exclusion of automated means for the removal of content. As a result, whereas there is no general obligation for providers to monitor proactively

[364] Section 116.

[365] Section 103(2)(b).

[366] Article 19 (2022), 'UK: Online Safety Bill is a serious threat to human rights online', https://www.article19.org/resources/uk-online-safety-bill-serious-threat-to-human-rights-online/.

[367] See for instance, Kettemann, M.C. (2019), 'Follow-Up to the Comparative Study On "Blocking, Filtering And Takedown Of Illegal Internet Content"', Leibniz-Institute for Media Research, Hans-Bredow-Institut, May 2019, p. 5.

[368] The law mentions that the operators of a digital platform have the duty to cooperate to combat disinformation, inter alia by the means of 'the fight against accounts that diffuse false information'.

generally, the choice of the specific measures to prevent the spread of terrorist content online is left to their discretion[369].

The CSAM proposal in particular creates such problematic general monitoring duties, which could significantly **affect the integrity of private online communications across the EU, set a negative example, and open the gateway for future general monitoring of content online**. Its rules essentially force providers to conduct **generalised monitoring of people's private communications**—even those that are encrypted[370]. The proposal requires such generalised monitoring to be done not only for verified illegal child sexual abuse material, which means that it will have been assessed by authorities to ensure that it is unlawful, but also for new photos and videos, as well as evidence of text-based 'grooming' both of which entail the recourse to AI-based scanning tools of potentially intimate conversations.

Legislation permitting or not expressly preventing generalised monitoring of online content based on the concern that it might be illegal or harmful can translate into forms of surveillance that are **in tension with the principles of legality, necessity and proportionality** which, under EU law and international and regional human rights law protect individuals **against arbitrary and unlawful restrictions to their rights to privacy and data protection.**

A line of CJEU jurisprudence has held that **online content monitoring requirements must be specific**. The Court of Luxembourg ruled that blocking a users' access to a particular website featuring content in breach of copyright law could be imposed by a platform, pursuant to a Court's order[371]. In cases that concerned the broader requirement of the implementation of a filtering system to prevent copyright infringements[372], the Court ruled that this type of requirement, whether imposed on an internet service provider (Scarlet) or a social network (Netlog), amounted to an obligation of general monitoring and was illegal pursuant Article 15 of the ECD.

More recently, however, the CJEU held that Facebook could be ordered by a national court to find and delete all comments identical to an original defamatory comment[373]. The Court held that a judicial order imposing a platform to remove content that is 'identical' or 'equivalent' to content previously ruled unlawful by a national court was permissible. Such an order can effectively require forms of proactive monitoring. For the platform to comply with an injunction to identify and remove content previously declared unlawful, it will in fact have to monitor all content which is identical or equivalent.

---

369 https://www.law.kuleuven.be/citip/blog/the-new-regulation-on-addressing-the-dissemination-of-terrorist-content-online/.

370 https://www.patrick-breyer.de/en/chat-control-leaked-commission-paper-eu-mass-surveillance-plans/EDRi

371 CJEU, 27 March 2014, UPC Telekabel Wien, C-314/12.

372 CJEU, 24 November 2011, Scarlet v. Sabam, C-70/10 and CJEU, 12 February 2012, Sabam v. Netlog, C-360/10.

373 CJEU, 3 October 2019, Glawischnig-Pieczek v. Facebook, C-18/18.

The Courts have attempted to circumscribe the scope of lawful exercise of online content monitoring duties to monitoring activities limited to a specific case, and associated to facts related to an individual infringement previously ascertained through independent judicial scrutiny[374]. At the same time, **the CJEU's decision to enable online content monitoring associated with a specific case has been subject to criticisms for its failure to clarify exactly what constitutes 'equivalent' content** that companies could be judicially mandated to take down, including without requirement of human intervention, via 'automated search tools and technologies'[375].

The DSA does not mandate general surveillance of online activities and maintains the non-liability rule of the e-Commerce Directive. However, **even voluntary measures can translate into a potentially unlawful interference with the rights to privacy and data protection**. This is so in the case of very large platforms and the measures they are expected to take with regard to systemic risks. Given the already endemic monitoring of individual behaviour, **these tasks may increase risks of more extensive monitoring and profiling of individuals.**

With a view to address these risks, **safeguards are necessary in the legal framework to ensure that there is an appropriate framing of the situation to ensure that monitoring of online content with certain systemic risks**— including through the use of automated data processing— **takes place in line with applicable privacy and data protection rules**. In its Opinion on the DSA proposal, the EDPS called for the inclusion of **provisions qualifying the types of illegal content that may warrant use of automated detection techniques involving the processing of personal data** and delineating the circumstances in which voluntary notification may take place[376]. However, these issues do not feature in the final text which does not address these issues. As will also be examined below in *subsection IV.1.3*, automated means may be used by online content moderation but the DSA does not impose specific limitations in that regard.

A final issue that needs to be flagged as well and concerns TERREG in particular related to the fact that in view of Europol's role in this respect, the oversight of its work and the applicability of different data protection rules/frameworks (GDPR and Europol Regulation) which apply will makes the work of EU level oversight or supervisory bodies more difficult in practice, as underlined by the EDPS regarding the latest Europol reform[377].

[374] Ibid., para 34.

[375] Ibid., para 47.

[376] EDPS, Opinion 1/2021 on the Proposal for a Digital Services Act (10 February 2021) para 27.

[377] See https://edps.europa.eu/press-publications/press-news/press-releases/2022/edps-orders-europol-erase-data-concerning_en.

IV.1.2. Exchanges of data between public authorities and online platforms

Privacy-related concerns have also been expressed in relation to initiatives enabling governments' proactive involvement in online content moderation[378]. **Government's monitoring of information that is publicly available about a person, such as social media posts, undoubtedly impinge upon the right to privacy**[379]. And yet, there is currently a wide array of EU and national agencies now policing the online environment through a variety of activities directed at identifying illegal or harmful content, and at enabling prosecution of crime[380].

Public authorities' engagement in this domain may take the form of **direct surveillance of information on online platforms through the probing of 'risk individual' or 'risk communities'** (e.g., extremist, conspiratorial, and/or religious groups) to tackle propagation of illegal or harmful content, including false information on social networks[381]. It also increasingly entails **governments access to and use of users data exchanged between online platforms and law enforcement authorities** on the basis of cooperation mechanisms established pursuant to laws designed to combat the production and dissemination of terrorist content, hate speech, and/or disinformation online.

For instance, concerns have been expressed with regard to the amendments made to the German *NetzDG* which require platforms to proactively provide data suspected to constitute criminal content to the federal police at the time it is reported by a platform's users[382]. This provision therefore enables law enforcement authorities' access to data, without prior court approval.

The need for *ex ante* independent scrutiny of law enforcement access to data for criminal investigation-related purposes is confirmed by a well-established body of jurisprudence by European courts. The CJEU has long established that in order to be lawful under EU law, requests for data for combating crime need **the prior validation of an independent administrative and/or judicial authority in the country of issue**[383].

Judicial oversight is required to verify that the gathering of electronic information can bring an effective contribution to the prosecution of a specific crime. This happens when in a specific case objective evidence is given that a relationship exists between the data and the person

[378] See, Report of the United Nations High Commissioner for Human Rights (2014), The Right to Privacy in the Digital Age, Human Rights Council, U.N. Doc. A/HRC/27/37, June 30, pp. 21-27.

[379] Ibid., p. 6.

[380] Europol (2020), 'Catching the Virus: Cybercrime, Disinformation And The Covid-19 Pandemic'.

[381] Jeangène Vilmer, J-B., Escorcia, A., Guillaume, M. and Herrera, J. (2018), 'Information Manipulation: A Challenge For Our Democracies'.

[382] Hardinghaus, A. Kimmich, R. and Schonhofen, S. (2020). 'German government introduces new bill to amend Germany's Hate Speech Act, establishing new requirements for social networks and video-sharing platforms', *Technology and Law Dispatch*, April 6.

[383] Joined cases C-293/12 and C-594/12 Digital Rights Ireland Ltd v. Ireland, 8 April 2014, para 62.

likely to be involved in the commitment of a crime[384]. The competent **law enforcement authorities are required to submit a 'reasoned request' from which it can be inferred that access to the data is strictly necessary for the purpose of prevention, detection or prosecution of crime**[385]. The reasoned request must be reviewed either by a court or by an independent authority prior to data access by the prosecuting authorities.

Similarly, the ECtHR determined that the acquisition by a public authority of communications data from a communications services provider requires that **data access be subject to prior review by a court or independent administrative body**[386]. Otherwise, the requirement derived from its own case law regarding the need for any interference with the rights of Article 8 of the ECHR to be 'in accordance with the law' cannot be regarded as fulfilled.

At the EU level, specific mechanisms for exchanges of data between law enforcement and the private sector have been developed. Article 14(5) of the TERREG in particular prescribes that where platform providers become aware of terrorist content involving an 'imminent threat to life', they must promptly inform authorities competent for the investigation and prosecution of criminal offences in the Member States concerned. Where it is impossible to identify the Member States concerned, the hosting service providers must notify the contact point in the Member State where they have their main establishment or where their legal representative resides or is established and transmit information concerning that terrorist content to Europol for appropriate follow-up. This rule may prove to be particularly difficult to implement and may result in that the reporting duties of providers are exercised frequently in absence of any supervision. What constitutes an imminent threat to life will depend on the interpretation given by each provider.

At the wider international level, **concerns have also been expressed about the possibility that initiatives such as the UN Cybercrime Convention** could be applied to enable law enforcement authorities cross-border access to data in the fight against content-based crimes such as hate speech, copyright infringement, and publishing disinformation[387]. These concerns relate to the possibility that access to law enforcement is too quick with too little due process guarantees[388].

---

[384] This happens when, in a specific case, objective evidence is given that a relationship exists between the data sought and the person likely to be involved in the commitment of a crime. Joined Cases C-203/15 and C-698/15 Tele2 Sverige AB v Post-och telestyrelsen and Secretary of State for the Home Department v Tom Watson, Peter Brice, Geoffrey Lewis.

[385] Joined Cases C-293/12 and C-594/12 Digital Rights Ireland Ltd v Ireland, 8 April 2014, para. 62.

[386] Judgment of the Court (First Section) of 13 September 2018, Case of Big Brother Watch and Others v the United Kingdom, App. No. 58170/13, 62322/14, 24960/15; see especially paras 463, 466, and 467.

[387] United Nations Human Rights Office of the High Commissioner (2022), OHCHR key-messages relating to a possible comprehensive International Convention on countering the use of Information and Communications Technologies for criminal purposes, 17 January 2022. See also Rodriguez, K and Gullo, K. (2022), 'Negotiations Over UN Cybercrime Treaty Under Way in New York, With EFF and Partners Urging Focus on Human Rights, Electronic Frontier Foundation', March 3.

[388] Ibid.

IV.1.3. Automated processing of wide range of sensitive personal information

Another specific data protection-related challenge arises from the fact that **online content moderation often requires the processing of a wide range of personal information**. Such processing may concern special categories of data, including information related, for instance, to political opinions, religious or other beliefs, trade union membership.

As regards TERREG, it is true that service providers are not obliged to use automated means; Article 5(2) vaguely refers to appropriate technical and operational measures or capacities, such as technical means to identify and expeditiously remove or disable access to terrorist content. Recital 24 states that where automated means are used, providers must indicate to the competent authority of the Member State whether they have the necessary capacity for human oversight and verification. Recital 25 further notes that if these are insufficient to address the risks, it may require the adoption of additional appropriate, effective and proportionate specific measures, which should still not be general monitoring or an obligation to use automated tools. However, it should be possible for hosting service providers to use automated tools if they consider this to be appropriate and necessary to effectively address the misuse of their services for the dissemination of terrorist content.

This wording could be read as **a broad encouragement for providers to resort to algorithmic systems**, which could mainly be used for two purposes; to prevent the further dissemination of digital objects that have been already labelled as terrorist content, and to identify new terrorist content[389]. Identifying new content based on machine learning systems will be able to sift through uploaded content to identify those objects which seem to be possible terrorist content, which is then prioritised by human reviewers.

As for the DSA, **the use of AI tools to proactively identify illegal or harmful content is subject to concretised and strict transparency requirements**. As noted in *Section II* of this Report, the terms and conditions will have to transparently set out information with regard to any restrictions imposed on the use of the service, including any algorithmic decision-making and human review (Article 12). In addition, providers will further have to publish reports on any content moderation they engaged in providing information also on any use made of automated means for the purpose of content moderation. In addition, the Commission as well as the Member States will have access to the algorithms of very large online platforms.

The use of automated tools is perhaps inevitable given both the sheer volume of online content, and the variety of conduct that may be removed (from sexual exploitation and hate speech to spam and bullying)[390]. Thus, **unsurprisingly platforms already utilise AI techniques to assist in**

[389] Bellanova (2022), op. cit., p. 11.

[390] For example, every minute 350 000 tweets are posted, 1 300 hours of video are uploaded to YouTube2 and, on Facebook, 510 000 comments are posted, 293 000 statuses are updated and 136 000 photos are uploaded. See Macdonald, S., Correia, S. and Watkin, A.-L. (2019), 'Regulating terrorist content on social media: automation and the rule of law', *International Journal of Law in Context*, Vol. 15, No 2, p. 184.

the enforcing the relevant rules. Overall, **recourse to AI systems must be made in a cautious and targeted way and following a risk assessment**.

As pointed out in *Section II*, the identification of terrorist content online may involve the identification of individuals and can thus amount to automated decision-making, including profiling, for example because of the requirement for preservation of data related to removed content based on Article 6. This will trigger the application of **Article 22 of the GDPR which prohibits a solely automated decision-making, producing legal effects on individuals**. Article 22(2) GDPR prescribes exceptions in this respect, particularly when authorised by EU or Member States and laying down 'suitable measures' to safeguard individuals' rights and freedoms as well as legitimate interests, namely provision of specific information to the data subject, the right to obtain human intervention, in order to express their point of view and to obtain an explanation of the decision reached after such assessment and to challenge the relevant decision[391].

In light of this, **human oversight and verification mechanisms should always take place as a safeguard that the decisions taken are accurate and well-founded**. Furthermore, **the increased transparency requirements in the DSA are a positive development**, as it will enable oversight on the algorithms developed and to information provided to the public about the use of algorithmic means.

IV.2. FREEDOM OF EXPRESSION

Freedom of expression and the right to information, including in the online environment, play a fundamental role in **facilitating and ensuring citizens' political participation and informed engagement in democratic decision-making processes and public debates**. They are in this way intrinsically embedded in **the notion and institution of democracy**[392]. Freedom of expression has been also understood as a prerequisite of other human rights, such as the freedom of opinion. Therefore, any limitations to this rights must remain exceptional and not undermine the very essence of the right itself[393].

A number of challenges emerge from the analysis of national and transnational initiatives related to the fight against illegal or harmful content online, pursuant to which restriction to protected speech are introduced and implemented in tensions with legality, necessity and proportionality requirements. These related the scope of the illegal content in online moderation (*Section IV.2.1*); the moderation of online content considered to be not illegal but 'harmful' (*Section IV.2.2*); and the use of automated means (*Section IV.2.3*).

---

[391] Recital 71 of the GDPR.

[392] J. Bayer et al. (2019), 'Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States', Study for the European Parliament, Brussels, Section 2.4.

[393] United Nations Special Rapporteurs Comment on 'A Counter-Terrorism Agenda for the EU: Anticipate, Prevent, Protect, Respond', COM(2020) 795, 21 October 2021, OL OTH (229.2021) (ohchr.org).

IV.2.1. The scope of the illegal content in online moderation

There are different content moderation challenges in addressing content framed as 'illegal'. Even within the wider category of illegal content, there are sub-categories, including content that is part of a wider offence and that has an online component (e.g., copyrights infringements, commercial scams and frauds), and content that is illegal per se, regardless of context (e.g., child sexual abuse material).

Online platforms are regularly required to restrict content qualified as illegal *per se* (e.g., representations of child sexual abuse, direct and credible threats of harm and incitement to violence). Under national and EU law, they are currently mandated to prevent posting or remove illegal content with substantial penalties for non-compliance within very short time frames (e.g., under the TERREG Regulation). However, **a key issue with regard to normative or policy initiatives** aimed at the fight against illegal content online **is that, often, they fail to provide an exact or sufficiently precise definition of the type of content that actually qualifies as 'illegal'**.

For instance, according to the 2016 EU Code of Conduct on countering 'illegal hate speech' online, the latter is to be understood by reference to the definition given to this term by Framework Decision 2008/913/JHA on combating certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law[394]. The scope of the restrictions that can be imposed pursuant to this Framework Decision, which also criminalises speech constituting 'incitement of hatred' (i.e., an emotional state or opinion), go well beyond those currently permissible under international law, under which restrictions to the freedom of expression should be limited to incitement of discrimination, hostility, or violence (i.e. threats linked to risks of specific actions). In addition, while the Framework Decision provides a lists of various types of proscribed conduct[395], it also allow Member States with a large margin of manoeuvre in the determination, through national transposition laws, of the severity threshold required to criminalise online speech[396].

The resulting picture is one where there are currently widely different legal approaches to the fight against illegal 'hate speech' across the EU which online platforms have been encouraged to enforce across their services[397]. Several Member States have adopted broadly worded restrictive laws criminalising various forms of 'extremism', 'blasphemy' and other instances of 'offensive speech'[398]. **Such vague definitions of illegal online content are at odds with the legality and legal certainty principles**, according to which restrictions to the freedom of speech must be 'provided by law' not only through regular legal processes, but also in ways which limit

[394] 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.

[395] Ibid. Art. 1(1).

[396] Ibid. Art. 7.

[397] Bukovská, B. (2019), 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online An analysis of freedom of expression implications', Transatlantic Working Group Paper, 7 May, p. 6.

[398] Venice Commission (2010), 'Blasphemy, insult and hatred: finding answers in a democratic society', *Science and technique of democracy*, No 4.

governments' discretion in a manner that distinguishes lawful and unlawful expression with sufficient precision.

The spreading of 'false news' is also considered illegal in an increasingly wide number of EU Countries, several of which expressly subject disinformation to criminal law and sanctions[399]. This is the case for instance in Austria[400], Croatia[401], Cyprus[402], Czech Republic[403], France[404], Greece[405], Italy[406], Malta[407], Lithuania[408], and Slovakia[409].

New laws criminalising the spread of disinformation have been also enacted in the wake of the Covid-19 pandemic. In Hungary, for instance, the Criminal Code's definition of 'scaremongering' was amended to include the dissemination of 'any untrue fact or any misrepresented true fact with regard to the public danger that is capable of causing disturbance or unrest in a larger group of persons at the site of public danger' or 'any untrue fact or any misrepresented true fact that is capable of hindering or preventing the efficiency of protection'[410]. The Covid-19 also prompted Bulgaria to adopt legislation criminalizing the spread of "internet misinformation" and granted the national media regulator the power to suspend websites for distributing disinformation[411].

The key risk in such cases is that the law serves as tool of online censorship through criminalisation directed at shaping the online regulatory environment and at suppressing legitimate discourse. The European Commission itself warned that criminal laws passed during the Covid-19 pandemic which introduce new crimes in relation to disinformation can lead to 'self-censorship' and raise 'particular concerns as regards freedom of expression'[412]. Focusing legislation related to online content moderation on criminalisation of disinformation can violate

[399] For a recent overview, see Ó Fathaigh, R. & Helberger, N. and Appelman, N. (2021), 'The perils of legally defining disinformation', *Internet Policy Review*, Vol. 10, No 4.

[400] Strafgesetzbuch [StGB] [Penal Code] § 264 [Verbreitung falscher Nachrichten bei einer Wahl oder Volksabstimmung] [Dissemination of false news in an election or referendum].

[401] Law on Misdemeanours against Public Order and Peace, Art. 16

[402] Criminal Code (Cyprus), Art. 50.

[403] Criminal Code (Czech Republic), Sec. 357

[404] France Freedom of Press Law, Art. 27.

[405] Criminal Code (Greece), Art. 191.

[406] Criminal Code (Italy), Art. 656, 658, 661.

[407] Criminal Code (Malta), Art. 82 (amended by the Media and Defamation Act, 2018).

[408] Law on the Provision of Information to the Public, No I-1418, art. 2(13) (1996), amended by No XII-2239 of Dec. 23 2015.

[409] Criminal Code (Slovak Republic), Sec. 361.

[410] Act XII of 2020 on containment of coronavirus.

[411] Organization for Security and Co-operation in Europe (OSCE) (2020), 'COVID-19 Response in Bulgaria Should Not Curb Media Freedom, Says OSCE Representative on Freedom of the Media', 15 April.

[412] European Commission (2020b), 'Communication on Tackling COVID-19 disinformation—Getting the facts right', JOIN/2020/8 final, p. 11.

the principle of proportionality. **The penalisation of disinformation has indeed been qualified as 'disproportionate' under international human rights law**[413].

Yet, certain European countries have enacted far-reaching measures in this regard. In Italy, for instance, the Ministry of Interior implemented an online reporting service known as the 'Red Button Protocol'. The Protocol enables users to report 'fake news' online to a specialised police force (the *Polizia Postale*) and has been referred to as a 'pipeline' for criminal prosecution with particularly concerning 'chilling effect' on the exercise of the right to freedom of expression[414]. These kinds of initiatives are difficult to reconcile with the UN Human Rights Committee's finding that prosecution for the 'crime of publication of false news' on the ground that the news was false, is in 'clear violation' of the right to freedom of expression[415].

The ECtHR has held that **prosecution for 'dissemination of false information', including under national election legislation, can violate the right to freedom of expression under Article 10 ECHR**[416]. That notwithstanding, France introduced a new law in 2018 giving the possibility for courts to order online platforms to remove 'inaccurate or misleading allegations or imputations of fact', which may 'alter the sincerity of an upcoming vote'. Similar legislation also exists in Poland, where courts may issue an order restraining the publication of 'untrue data or information' about an election candidate[417]. The ECtHR found Polish judicial proceedings for 'untrue information' during an election to suffer from an overall lack of fairness[418], which translates into a violation of Article 10 of the Convention. The Polish law was also found to violate Article 10 because the large discretion it entrusts to national courts to qualify statements as 'lies' deprived concerned individuals (i.e., a politician) of the protection granted under such provisions[419]. In a subsequent case, the ECtHR established that Polish courts 'unreservedly qualified all of [the statements] as statements which lacked any factual basis[420].

With regard to TERREG, terrorist content concerns different types of material that: (a) incites the commission of one of the terrorist offences as defined in Directive (EU) 2017/541[421], where such material, 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; (b) solicits a person or a group of persons to commit or contribute to the

[413] OSCE (2020), 'Joint Declaration on Freedom of Expression and Elections in the Digital Age', para 42.

[414] David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. OL ITA 1/2018, at p. 1, p. 2.

[415] Human Rights Committee, 1999, para. 24.

[416] Salov v. Ukraine, 2005, para. 113.

[417] Law of 16 July 1998 on Elections to Municipalities, District Councils and Regional Assemblies, Dz.U. 1998 Nr 95 poz. 602, para. 72.

[418] ECtHR, Kwiecień v. Poland, 2007, para. 55

[419] ECtHR, Brzeziński v. Poland, 2019, para. 58.

[420] ECtHR, Kita v. Poland, 2008, para. 51.

[421] Except threatening to commit one of the offences.

commission of one of the offences; (c) solicits a person or a group of persons to participate in the activities of a terrorist group; (d) provides instruction on the making or use of explosives, firearms or other weapons or noxious or hazardous substances, or on other specific methods or techniques for the purpose of committing or contributing to the commission of one of the terrorist offences; (e) constitutes a threat to commit a terrorist offence[422].

Arguably, the final text is marginally better than the Commission proposal which referred to 'encouraging' the contribution, participation, or support to terrorism or a terrorist group. That term could certainly encompass legitimate forms of expression, such as reporting conducted by journalists and human rights organisations on the activities of terrorist groups and on counter-terrorism measures taken by national authorities. The United Nations' Special Rapporteurs on human rights have emphasized that **TERREG includes an overly broad definition of terrorist content that may encompass legitimate expression protected under international human rights law'**[423]. The final text is far clearer in that respect: Recital 12 excludes material disseminated for educational, journalistic, artistic, or research purposes or for awareness-raising purposes against terrorist activity for the scope of TERREG.

The reference in the proposal to 'material that incite or advocates committing terrorist offences, promotes the activities of a terrorist group or provides instructions and techniques for committing terrorist offences has also been criticised as too broad[424]. However, **the addition in the finally adopted text (compared to the proposal) that terrorist content can involve a 'threat' to commit a terrorist offence seems rather broad and has the potential to become a residual category**, as opposed to the more precise wording of the other kinds of material that constitute terrorism content.

As noted in *Section II* of this Report, the DSA defines illegal content by reference to what is considered in itself or in relation to an activity not in compliance with either EU law or the law of *any* Member State, irrespective of the precise subject matter or nature of that law. Recital 12 explains that the concept of illegal content should 'broadly reflect the existing rules in the offline environment; in particular, the concept should be defined broadly to cover information relating to illegal content. In particular, that concept should be understood to refer to information, irrespective of its form, that under the applicable law is either itself illegal, for example, illegal hate speech or terrorist content and unlawful discriminatory content, or that the applicable rules make illegal in view of the fact that it relates to activities that are illegal.

The rather broad definition of illegal content entails that the DSA does not impose any limits as to what content can be criminalised at the national level and the clarifications do not help in this regard. The legality of national criminal laws is presumed and may be the basis for blocking

[422] Article 2)7) of the TERREG.

[423] Kaye, D., Cannataci, J. and Ní Aoláin, F. (2018) Letter to the EU [doc. OL OTH 71/2018], Geneva: Special Procedures of the UN Human Rights Council, p. 2.

[424] https://www.eff.org/el/deeplinks/2021/04/eu-online-terrorism-regulation-bad-deal?fbclid=IwAR2f7CWZaDBfYKE_zqXCKdul2HGZ41GDNwmtDL0nqsEcV3dAxv-rC4CmBKI

content which Is considered as illegal even though the law that considers it so does not comply with the principle of legality. As a result, **problematic and unlawful criminal laws at national level will continue to be the basis of online content moderation impinging up the freedom of expression of users**.

The Online Safety Bill follows a somewhat different approach. Illegal content is content that amounts to a relevant offence. The latter were originally oriented towards terrorism offences and child sexual abuse. In its latest iteration of the Bill, other specific offences are written on the face of the bill as priority offences. These are: revenge porn, hate crime, fraud, the sale of illegal drugs or weapons, the promotion or facilitation of suicide, people smuggling and sexual exploitation. Naming these offences removes the need for them to be set out in secondary legislation at a later stage and possibly be expanded in the future. Furthermore, Ofcom can take faster enforcement action against tech firms which fail to remove the named illegal content[425]. **As long as the listed offences are targeted and limited to what is strictly necessary and framed under clear terms, the UK approach appears more balanced than the one outlined in the DSA**.

That said, the Bill exempts news publisher content, which involves content generated by a 'recognised news publisher' or content that reproduces or links to the full version of an article originally published by a recognised news publisher[426]. According to the Bill, a recognised news publisher must hold a broadcasting licence and publish news-related material in connection with the activities authorised under the licence. Alternatively, a publisher must meet a series of criteria, such as having a business address in the UK and publishing news-related material (subject to editorial control and in accordance with a standards code) as its principal purpose[427]. As a result, foreign news publishers will likely not be able to avail themselves of that exemption[428]. In addition, citizen journalists may not be able to fulfil the exemption criteria, even though they may engage in vital journalistic activity[429].

IV.2.2. The moderation of harmful online content

Increasingly, online content moderation initiatives are aimed at tackling the dissemination of online content that is not illegal, but that is labelled as 'harmful'. The EU Code of Conduct, for instance, introduced content-based restrictions also targeting content that is not explicitly illegal, in particular encouraging platforms to also prohibit 'hateful conduct.' Task Force discussions confirmed that **this is a vague term that is not limited to illegal content** (i.e., content expressly criminalised by law at the national or EU level) **and that can, potentially, encompass mere vulgar speech.**

---

[425] https://www.gov.uk/government/news/online-safety-law-to-be-strengthened-to-stamp-out-illegal-content

[426] Clause 40.

[427] Section 50.

[428] Article 19.

[429] Ibid.

However, speech which is offensive, shocking, or disturbing might be permitted. Under European human rights law, in particular, the threshold to be met in order to justify restrictions to such freedom is very high, since protection under Article 10 ECHR extends to material that can 'offend, shock or disturb'[430]. The ECtHR has clearly established that the offensive character of speech does not automatically deprive it of its status as expression falling under Article 10.

Regarding the specific categories or types of hate speech that are not worthy of protection, there is not one single definition. Different human rights instruments provide different definitions, and the ECtHR has so far adopted a case-by-case approach, avoiding engaging in value judgments when deciding if certain material labelled as hate speech deserves protection. The rationale underpinning this approach seems that of avoiding a slippery slope towards greater censorship based on a majoritarian view of what is or is not 'valuable' speech. So far, the qualification of hate speech has been reserved to particularly loaded manifestations (e.g., promotion and justification of terrorism and war crime, incitement to violence, negation of holocaust, promotion of totalitarian ideologies but also threat to constitutional orders, which include speech that engenders operation of free and democratic institutions)[431]. However, **when the term hate speech is generally referred to 'hateful conduct', it can assume a wider range of meanings.**

The ECtHR has confirmed that is not reasonable to restrict freedom of expression only to generally accepted ideas. In case of deliberate disinformation, that the Commission defined 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'[432], is not speech particularly worth of protection. This means that states and companies implementing content moderation policies at the national or transnational level might have a wider margin of discretion. At the same time, **definitions of disinformation are too broad and vague from the perspective of legal certainty, effectiveness, and freedom of expression.**

Serious legal certainty challenges arise in a context where **there is a lack of common understanding of what exactly constitutes harmful (but not illegal) online content,** and there is a profound unclarity about **what is the exact legal basis** for the adoption of content-based restrictions targeting the manifestations of online speech which do not constitute direct violations of the law. Of particular concern in this regard is **the enforcement by online platforms of restrictions of content that, while not illegal, is considered harmful or in some way unacceptable on the basis of companies' ToS or community guidelines**.

While this kind of content moderation may be justified on the basis of contractual breaches (i.e., violations of the ToS, or of Community standards), it cannot translate into undue

[430] ECtHR, *Handyside v. the United Kingdom*, para 49; ECtHR, Observer and Guardian v. the United Kingdom, para 59.

[431] ECtHR (2021), Guide on Article 10 of the European Convention on Human Rights – Freedom of Expression, Updated 30 April 2021.

[432] European Commission (2018), Tackling Online Disinformation: A European Approach COM/2018/236 final, Section 2.1.

interferences with fundamental rights and freedoms. **Governments and regulatory authorities at the national and supranational level not only have the obligation to refrain from unduly interfere with freedom of expressions. They also have positive obligations to protect such freedom, including from private individuals' interferences.** This means that protective measures are needed to guarantee individuals against the ways in which these powers are exercised.

The scope of the DSA aims to address the societal risks that the dissemination of disinformation or other content may generate[433]. Providers must pay attention how their services are used to disseminate or amplify misleading or deceptive content, including disinformation with a real and foreseeable negative impact on public health, public security, civil discourse, political participation and equality[434]. The DSA allows providers to make of trusted flaggers or similar mechanisms to take quick and reliable action against content that is incompatible with their terms and conditions, in particular against content that is harmful for vulnerable recipients of the service, such as minors[435]. However, contrary to the concept of 'illegal content', **the DSA does not define what content qualifies as 'harmful'**, through there are indications throughout the text that disinformation with an impact on public health, or that affects minors is of special interest.

The Online Safety Bill also imposes an obligation on large platforms to take down and restrict access to content that is entirely legal but considered 'harmful'. This is notwithstanding strong opposition by civil society and the explicit recommendation of the Joint Committee on the Draft Online Safety Bill to steer away from the concept of 'legal but harmful' content. This requirement will essentially require platforms to come up with terms of service to restrict such content 'inevitably err[ing] on the side of censorship'[436]. As a result, new categories of speech are created which are 'legal to say, but illegal to type'[437]. Legal speech is protected speech and legislation requiring online platforms to censor legal speech does not comply with international standards on freedom of expression.

Furthermore, **the definition as to what constitutes 'content that is harmful to adults' does not meet the legality requirement under international human rights law, as it is too vague and unclear**. Section 54 includes content 'of a kind which presents a material risk of significant harm to an appreciable number of adults in the United Kingdom'. The inclusion of a potentially objective standard that the adult must be of ordinary sensitivities was removed from the draft version of the Bill. Their further specification will be laid down in secondary legislation following consultation with Ofcom, and that legislation may be subject to constant revision. This is to

[433] Recital 9.

[434] Recital 57, 63.

[435] Recital 46.

[436] https://www.theguardian.com/technology/2022/jul/13/online-safety-bill-tories-free-speech-david-davis.

[437] Ibid.

ensure most recent evidence and emerging harms can be added quickly, future-proofing the legislation[438].

Nevertheless, these definitions as they currently stand do not provide much legal certainty to individuals as to what content will be moderated and therefore what they can or cannot write online particularly to individuals they do not know. Due to the heavy fines, it is highly likely that providers will not risk being penalised and may have a sweeping approach of over-removals of legal content, if there is even the remote chance that it could be considered harmful to adults with subjective sensitivities[439]. As Trengove et al. have stressed, '[w]hether an individual piece of content counts as 'harmful' on this definition will depend significantly on the context of its use. It is this matter of interpretation that, we think, opens the possibility of over-censorship when enforced by a top-down approach in which the government specifies a list of harmful content or algorithmic systems are used to detect harmful content'[440].

As a response, the UK government has published an indicative list of priorities as regards what constitutes harmful content for children and adults[441]. In relation to adults the list concerns: online abuse and harassment (but mere disagreement would not reach the threshold of harmful content); circulation of real or manufactured intimate images without the subject's consent; content promoting self-harm, eating disorders or on legal suicide; harmful that is demonstrably false, such as urging people to drink bleach to cure cancer or vaccine disinformation. For children the harmful content is divided in two categories: those where children must be prevented from encountering altogether (pornography, self-harm or eating disorder promoting content, legal suicide content, and content where providers must ensure age appropriateness (online abuse, cyberbullying and harassment, harmful health content, and content depicting or encouraging violence.

The aim of this list is to ensure that a distinction is drawn between strongly felt debates on the one hand, and unacceptable acts of abuse, intimidation, and violence on the other. It is a step forward, focusing on a risk-based approach, thus on the bigger risks of harm. However, it is indicative and nothing prevents the government from expanding it. Furthermore, it includes notions which are broad and whose exact scope and reach is not fully clear. Furthermore, it is difficult to grasp how providers will determine whether the threshold for harassment for example has been met, so that they do not impinge upon freedom of expression and create chilling effect on users online.

---

[438] https://questions-statements.parliament.uk/written-statements/detail/2022-07-07/hcws194, 7 July 2022.

[439] Article 19.

[440] Trengove, et al., A critical review of the Online Safety Bill https://reader.elsevier.com/reader/sd/pii/S2666389922001477?token=744DAF8E41BC9BAB7BAA26012BFA41E5040427E68CFBEB496BAF65F9479AD1B81F3B346E3B48BFB249E86ABA7DB938C1&originRegion=eu-west-1&originCreation=20220731155522, p. 7.

[441] https://questions-statements.parliament.uk/written-statements/detail/2022-07-07/hcws194.

IV.2.3. The use of automated means to moderate online content

Beyond the data protection-related concerns discussed earlier, **the use of AI-based tools is problematic for other reasons**: the potential of general monitoring of the online environment by providers, coupled with the imposition of fines for non-compliance, may result in **creating incentives for over-removal of content**. In practice, providers would rather prefer to delete more content, faster, for example by installing uploading filters that will increasingly rely on automated means.

A key issue here is that recourse **to AI tools has been promoted as highly effective and reliable, but this is always not entirely true** for two main reasons: First, machine learning systems for detection and identification of potentially illegal content are **context blind and commit errors**, thus are not currently reliable. For example, owing to the lack of precision in making a distinction terrorist content from content related to, for example, activism or satire about terrorism, the use of automated tools could result in the inadvertent removal of lawful content[442].

Unless there is strict human review, the increasing reliance on AI-based tools will entail in that a number of erroneously flagged content will be removed infringing the freedom of expression. The impact on individuals may well go much further than violations of freedom of expression. For example, a picture being automatically flagged on social media because an AI tool wrongly considered that the picture contained nudity has far less serious consequences compared to law enforcement authorities accusing an individual of disseminating illegal child sexual abuse material or grooming a child.

Second, **different moderation tools tend to get fused and mixed under the umbrella of AI.** While some platforms providers are using machine learning techniques to identify new instances of harassment, hate speech, or pornography, most of them are resorting to 'pattern matching', which entails comparing new content to a blacklist of known examples. While this does not qualify as AI, it is still considered an automated means[443]. In other words, the content moderated that is automatically identified are copies of content that have already been reviewed by a human moderator. This signifies that claims of successful moderation based on AI tools are exaggerated, because in certain cases it is not AI tools that have led to successful moderation but other automated means[444].

---

[442] https://www.law.kuleuven.be/citip/blog/the-new-regulation-on-addressing-the-dissemination-of-terrorist-content-online/.

[443] Gorwa et al., 2020.

[444] Gillespie, T., (2020), 'Content moderation, AI, and the question of scale', *Big Data & Society*, vol. 7, No 2.

IV.3. RULE OF LAW, DUE PROCESS AND EFFECTIVE REMEDIES

Online moderation activities are increasingly enforced outside formal legal processes, pursuant to platforms' own terms of services, **often without any independent judicial oversight and supervision, which constitute central tenets in the notion of the rule of law. These are of crucial importance in delivering access to justice and remedies sufficient to ensure effective legal protection**.

The relevance of the principle of separation of powers and effective judicial protection by independent courts has been considered as a central tenet of the Union's notion of the rule of law. According to 2019 European Commission Communication titled 'Strengthening the rule of law within the Union: A blueprint for action', 'Under the rule of law, all public powers always act within the constraints set out by law, in accordance with the values of democracy and fundamental rights, and under the control of independent and impartial courts'[445]. The Luxembourg Court has concluded that 'the very existence of **effective judicial review** designed to ensure compliance with EU law **is of the essence of the rule of law**'[446].

The EU has developed several initiatives that promote **extra-legal content moderation policies by online platforms providers**. The EU IRU, for instance, enable law enforcement authorities to flag 'terrorist and violent extremist content' online and to cooperate with service providers with the aim of removing this content. The EU Code of Conduct on countering 'illegal hate speech' online also involves with **major online platforms to directly remove content**. The Code of Conduct led to the implementation of online content moderation activities by platforms pursuant to secretly adopted restrictions, including for instance those implemented the context of the Global Internet Forum to Counter Terrorism through the use of databases of hashes, which are particularly problematic from a legality and due process perspective.

Similar remarks apply to **the TERREG, which has been heavily criticised about the notorious one-hour deadline for online platforms to comply with orders for removal**. In essence, this deadline does not enable platforms to review the orders, is too rigid and simplistic and fails to address the concerns that platforms are becoming the arbiters of online speech and not the independent judiciary. **The requirement to remove allegedly illegal content in such a short time frame without the possibility to turn to the court empowers platforms with extra power without proper oversight**. Importantly, whereas competent authorities must carry out their

[445] European Commission (2020), Communication 'Commission Work Programme 2020: A Union that Strives for More, COM(2020) 37 final, 29 January. Furthermore, according to the Commission 2014 Communication on 'A New EU Framework to Strengthen the Rule of Law', there are certain 'shared principles' which lay at the core of the rule of law as a 'common value' in the Union, and which include the principle of legality (a transparent, accountable, and democratic process for enacting laws), prohibition of arbitrariness, independent courts, effective judicial review, and the respect for fundamental rights. European Commission (2014), Communication, A New EU Framework to Strengthen the Rule of Law, COM(2014)158 final, 11.3.2014.

[446] Refer to Court of Justice of the European Union (CJEU) case law: Case C-64/16, Associação Sindical dos Juízes Portugueses v. Tribunal de Contas; Case C- 216/18 PPU, LM, case C-619/18, Commission v. Poland (order of 17 December 2018); Order of the Court, Case C-441/17, European Commission v Poland, 20 November 2017; and Order of the Court, Case C-791/19, European Commission v Poland, 8 April 2020.

duties in an objective and non-discriminatory manner, the issue of judicial review of the decisions on the removal and thus of effective oversight has been left unanswered; neither the proposal nor the final text includes anything in that regard.

Complex questions of fact and law—including who monitors democracy, the rule of law and human rights— require to be adjudicated by competent public institutions, not private companies. Platform services actions are principally informed by economic interests, and whose internal processes cannot adequately satisfy due process standards. Recent research compared the length of judicial proceedings cases aimed at assessing the illegal or harmful nature of certain speech in selected Council of Europe States, with the timeframe within which some governments (and now the EU) require platforms to decide and take down illegal content pursuant to mandated notice and take down regimes. The analysis suggests that large discrepancies may lead platform services to err on the side of removal, rather than protecting rights and liberties of their users against (potentially censorious) governments[447].

In such context, online platforms providers— acting under the threats of financial penalties— become responsible to shape *the scope of protection of rights* by enforcing legal restrictions, as well as abstract notions such online harm, through digital surveillance and automated decision-making tools. Since artificial intelligence technologies are always becoming more pervasive in online content moderation, concepts of legality, necessity, and proportionality are no longer applied by competent oversight authorities, but increasingly defined through algorithmic calculations. This situation leads to the 'mathematisation of the law' which, given the opacity of the technologies deployed to perform content moderation duties, raises concerns not only for human rights protections, but more generally for the rule of law and democracy.

The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression recommended that platforms must only be required to remove content pursuant to an order issued by an independent and impartial judicial authority determining the unlawful nature of the relevant content[448]. The judicial nature of such orders is essential to give due and proper weight to content-based restrictions affecting fundamental rights including *inter alia* privacy, data protection and the freedom of expression.

Due process challenges also arise in relation to content moderation actions and procedures that not adequately reflect principles of fairness vis-à-vis concerned individuals, in particular those who do not agree with a platform's decision. To date, all major platforms have established in-house appeal mechanisms though which users can challenge decisions resulting, for instance, in removal of content, or account restrictions. However, the Task Force meetings showed that much discretion is still left to companies with regard to the level of independence, accessibility, transparency, and predictability in the context of internal oversight mechanisms

[447] https://futurefreespeech.com/wp-content/uploads/2021/01/FFS_Rushing-to-Judgment-3.pdf.

[448] UN Special Rapporteur (2018), 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression'.

and review procedures. Only in presence of adequate internal oversight infrastructures and complaint procedures individuals can be precisely informed of the basis upon which their content was removed, or of the reasons why their complaint did not lead to content being removed. These, in turns, are crucial preconditions for the exercise by affected individuals of the right to seek redress and obtain remedies.

The possibility for users affected by private actors' content moderation to access conventional judicial remedies is crucial, since it enables an independent scrutiny of the legality and fairness of actions taken pursuant to platforms' ToS. Ultimately, it is the Courts' responsibility to assess not only whether certain content/behaviours violated the law or are in breach of ToS or the Community Standards, but also to consider whether ToS and Community Standards (and content moderation actions taken pursuant to such documents) are legally valid[449].

Judicial redress against platform providers should be granted to users affected by the handling of individual content. Ultimately, the judicial system should act as an arbiter in disputed cases instead of government agencies or platforms.

This means that in addition to refraining from interfering with individuals' human rights, states are expected to: a) pass necessary measures enabling individuals to practically exercise their rights, and b) take steps to protect human rights if they are violated by another individual (or authority). The latter brings us to the 'horizontal effect' of human rights, in other words, the Convention's scope to relationships between private individuals. Horizontal effect means that private actors have to respect the fundamental rights of each other. Ultimately, this obligation can be enforced by the courts. This is therefore a triangular relationship between the state and two private actors.

[449] Holznagel, D. (2021), 'Enforcing the Rule of Law in Online Content Moderation: How European High Court decisions might invite reinterpretation of CDA', *Business of Law: Internet Law and Cyber-Security*, December 19.

SECTION V. CONCLUSIONS AND RECOMMENDATIONS: A PRINCIPLED AND RIGHTS-CENTRED LEVEL PLAYING FIELD FOR AN OPEN AND SECURE ONLINE ENVIRONMENT

Based on the key findings and analysis provided in this Report, this Section concludes by highlighting a set of key policy suggestions and recommendations. They focus on the need to developing a common level playing field on online content moderation in the EU, the UK, and globally which is firmly anchored and driven by a principled rule of law and human rights-centred approach.

Policy Recommendation 1: Ensuring regulatory clarity and quality and upholding the principle of legality

- The principles of legality and legal certainty require restrictions on online content to be clearly defined in law and subject to effective remedies. Such principle applies to prohibitions of content, but also to the introduction by states and at the international/regional level of content moderation obligations, and liability rules for non-compliance of online platforms.
- Tensions with the legality and legal certainty principles arise when the determination of the types of content to be considered illegal and harmful in nature depends upon loosely demarcated risks of 'online harm' or poorly defined categories of unlawful and harmful content or speech ('extremist', 'hate speech', or 'disinformation'). The lack of regulatory clarity generates the risk that platform services policies governing content moderation become vague, overinclusive, or give rise to discriminatory effects and undue restrictions of fundamental rights, democracy, and the rule of law.
- Precise rules and definitions must be provided by law to avoid incentives for online platform providers to impose unlawful and disproportionate restrictions by the means of their ToS and/or community standards. In order to prevent global standards on online content to be established *de facto* by online platforms, laws and regulations need to strictly define and limit categories of illegal content, including what qualifies as unlawful speech, hate speech, advocacy of terrorism, and harassment. This includes EU and national measures referring to illegal and harmful content, which should be reviewed to ensure that the definitions meet the criteria of legality and legal certainty.
- Broadly worded restrictive laws criminalising various forms of online 'extremism', 'blasphemy' and other instances of 'offensive speech' are at odds with the legality requirement, according to which restrictions to the freedom of speech must be 'provided by law' not only through regular legal processes, but also in ways which limit governments' discretion. This must be done in a manner that distinguishes lawful and unlawful expression with sufficient precision. Furthermore, criminalisation of content and speech should remain the regulatory option of last resort, and strictly limited to restrictions of content the production or dissemination of which negatively impact expressly recognised rights.

- Restrictions must be directed at protecting the legitimate interests at stake and impose the least burden on the exercise of the freedoms affected by content moderation restrictions. General monitoring of online content by online platforms should be explicitly banned. Content moderation duties should only be deemed permissible if they are *specific* in respect of both the protected subject matter and the potential infringers.

Policy recommendation 2: Ensuring accessible, foreseeable, and transparent online platform services' monitoring and oversight policies

- Judicial authorities and not administrative authorities— except in relation to online market places— should have the power to decide on the legality of content. This has not been achieved, since the co-legislator decided to rely on the model already adopted in the E-Commerce Directive, which recognised a role for administrative authorities. This raises challenges from the perspective of effective protection of the right to freedom of expression.
- Content moderation should take place in accordance with the rule of law. The involvement of judicial authorities is an essential precondition to ensure that restrictions on fundamental rights and freedoms remain as targeted as possible, in accordance with principles of necessity, proportionality, and data minimisation. The kind of networked and multilevel regulatory oversight and external accountability infrastructure established by the GDPR, under which DPAs' regulatory oversight and external accountability role complements, and is complemented by, judicial oversight mechanisms and related remedies, should also be extended to other EU policy areas or legal domains under which content moderation takes place.
- The complexity of online platforms content moderation infrastructures and processes makes content regulation incomprehensible for users. To ensure legitimacy, trust and effectiveness, the internal (i.e., non-administrative and non-judicial) accountability mechanisms for users to report harms and raise concerns related inter alia to freedom of expressions and privacy should be clear, accessible, predictable, and transparent. These standards should apply to the mechanisms' structure, mandate, procedural settings, and enforcement options.
- Content whose illegality has not been demonstrated should remain online. Online platforms are now mandated to create an internal complaint-handling system to enable individuals whose information has been affected by certain content moderation decisions to lodge complaints within a given time period. Following the complaint, the online platform must review the decision, and potentially reverse it if the content is found to be legitimate.
- Online platforms should be required to explain how they reach content moderation decisions. When resorting to deletion of content decisions, they should be required to provide the user with an explanation. In a context where AI technologies become more pervasive in online content moderation, and material application of legal rules and values depends on automated decisions making tools developed or used by online platforms, far-reaching fundamental rights, rule of law, and democratic challenges arise.

- Online platforms should be subject to strict transparency requirements extending *ex ante* to the design, prototyping and standardization, and testing of practical deployment of AI tools and procedures followed. To implement legally mandated online content moderation tasks AI should be tested and transparent and should not be used for experimentation on masses on people. AI tools for the performance of online content moderation should be classified as 'high risk'. This is not reflected in the proposal for an AI Act, which provides a classification of high risk systems in its Annex III, but no listed systems concern online content moderation. Concerned individuals should be granted the right to obtain human intervention on the part of the online platforms (acting as data processor or data controller), as well as the right to contest the decision. Recourse to AI systems must be made in a cautious and targeted way and following a risk assessment.
- Content moderation should, by design and by default, not involve the processing, collection, and disclosure of personal data to as great extent as possible. Furthermore, safeguards are necessary in the legal framework to ensure that there is an appropriate framing of the situation to ensure that monitoring of online content/certain systemic risks— including through the use of automated data processing— takes place in line with applicable privacy and data protection rules.
- Online platforms could be required to employ Freedom of Expression Officers (FEOs) specialised on regional and national legal standards and jurisprudence and trained to understand the national contexts of speech. Similar to the Data Protection Officers established under the GDPR, Freedom of Expression Officers could be tasked with the duty to oversee compliance of platforms' decisions on content moderation.
- Procedural fairness safeguards should be guaranteed not only in relation to how decision are taken, but also to ensure effective remedies to individuals. There should always be ways of rectifying wrong decisions. Minimum common standards in terms of online platforms internal oversight, and review mechanisms for content moderation decisions are therefore needed. Company's internal accountability procedures should not only ensure minimum levels of independence, but also be clear, accessible, predictable, and transparent.
- Platforms should only be required to remove content pursuant to an order issued by an independent and impartial judicial authority determining the unlawful nature of the relevant content. *Ex ante* judicial oversight should also be guaranteed in relation to the inclusion of certain content in databases of hashes. Regulators at the EU and national level could create specialized independent judicial bodies able to issue such orders in an expedited while still fully fair manner, while preserving core aspects of due process and attaching due and proper weight to the freedom of expression, in accordance with international and regional human rights standards.
- *Ex post* reporting duties should also be introduced to increase publicity and accountability of platform providers with regard to the online content moderation activities implemented, including types and numbers of online content restrictions adopted.

Policy Recommendation 3: Guaranteeing effective regulatory and networked oversight and enhancing multi-actor coordination

- Effective monitoring of online content moderation requires the regulation of duties and powers of independent administrative authorities entrusted with the responsibility to oversee that (in the implementation of online content moderation regulations) platforms comply with applicable laws and policies. Regulatory oversight cannot be limited to the field of data protection. To be effective, regulatory oversight must encompass a varied of specialized bodies covering multi-level normative and policy framework concerned, and encompassing provision of services in the digital marketplace, media laws, copyright protection, consumer protection, and competition law.
- Individuals confronted with an infringements of norms regulating online content moderation should have the right to lodge a complaint before the competent national regulatory oversight authority. This should be without prejudice to have the right to turn directly to the judiciary. Oversight authorities should be tasked with the responsibility to facilitate the submission of complaints, and with the duty to handle lodged complaints and with investigating, to extent appropriate, the complaints' subject matter.
- Concerned individuals should have the right to an effective remedy against the decisions of the authority, as well as in case of lack of action or lack of information about the progress or outcome of their complaints. It should be considered the possibility to allow complaints to be handled by multiple oversight authorities, cooperating through a sort of one-stop-shop mechanism. The possibility to enable a complaint to be lodged by Civil Society Actors, on behalf of the concerned individuals, should be considered.
- In light of the experience gained through the UK Digital Regulation Cooperation Forum, ensuring and supporting enhanced coordination and the sharing of information and 'promising practices' between different national oversight actors across the EU—as well as relevant EU agencies— can prove to be especially crucial in fostering more effective and consistent ways of digital area collaboration and the development of a level playing field on online content moderation. The newly envisaged European Board for Digital Services by the DSA is a welcomed initiative in this respect and may prove to play a crucial role in further developing Union expertise and capabilities. The impartiality and effectiveness of its role must be carefully monitored to ensure that it delivers its full potential.
- To ensure the effective accountability of online platforms, regulatory oversight actors must be endowed with additional financial resources as well as effective investigative and enforcement powers, including the power to impose corrective measures and financial sanctions in case where deviations from regulatory duties are found and verified. At the same time, administrative authorities should not have the power to issue decisions having the effect of infringing freedom of expression, without *ex ante* authorisation by an independent judicial authority.

ANNEX 1. TASK FORCE MEMBERS AND PARTICIPANTS

Jeff Alford
Security and Online Harms Policy Adviser, EU lead, Department for Digital Culture Media and Sport

Philipp Amann
Head of Expertise and stakeholder management, Europol

Cecilia Alvarez
Director of Privacy Policy Engagement, EMEA, Facebook

Asha Allen
Advocacy Director for Europe, Online Expression & Civic Space, Center for Democracy & Technology (CDT)

Judit Bayer
Senior Research Fellow at ITM, University of Münster, Germany

Andrea Beccalli
Stakeholder Engagement Senior Director-Europe, ICANN

Charlotte Brennan
Deputy Bill Manager for the Online Safety Bill, Department for Digital Culture Media and Sport

James Dipple-Johnstone
Deputy Commissioner (CRO) of the Information Commissioner's Office (ICO)

Sybe A. De Vries
Professor of Public Economic Law, Jean Monnet Chair, Utrecht Centre for Regulation and Enforcement in Europe (Renforce), Utrecht University

Adriana Edmeades Jones
Legal Advisor to the United Nations Special Rapporteur on Protection of Human Rights while Countering Crime and Terrorism

Marie-Louise Gaechter
Head of the Data Protection Authority of Liechtenstein, European Data Protection Board (EDPB) Representative

Catherine Garcia-van Hoogstraten
Director European Government Affairs, Microsoft

Raphaël Gellert
Assistant Professor, ICT & private law, Radboud University

Catalina Goanta
Associate Professor, Faculty of Law, Utrecht University

Balazs Gymesi
Communications Officer, OECD DisMis project

Hielke Hijmans
Chairman of the Litigation chamber and member of the Board of Directors of the Belgian DPA

Chris Jones
Executive Director, Statewatch

Alice Knight
International Regulation and Trade Policy (Online Harms), UK

Aleksandra Kuczerawy
FWO Postdoctoral researcher, Center for IT & IP Law, KU Leuven

Dario La Nasa
Senior Public Policy Manager, Twitter, Brussels office

Eva Lachnit
Senior Policy Advisor International Affairs, Dutch Data Protection Authority (Dutch DPA)

Clementina Salvi
PhD Candidate Queen Mary, University of London

Rowena Schoo
International Policy Manager, Ofcom, UK

Eva Simon
Senior Advocacy Officer, Civil Liberties Union for Europe

Brendan Van Alsenoy
Deputy Head of Unit "Policy and Consultation", European Data Protection Supervisor

Santiago Wortman Jofré
Policy Analyst Consultant, OECD

CEPS
PLACE DU CONGRES 1
B-1000 BRUSSELS