Parliamentary Assembly
Assemblée parlementaire
http://assembly.coe.int

COUNCIL OF EUROPE
75
1949 - 2024
CONSEIL DE L'EUROPE

Doc. …
Provisional version
4 December 2024

Regulating content moderation on social media to safeguard freedom of expression

Report[1]
Committee on Culture, Science, Education and Media
Rapporteur: Ms Valentina GRIPPO, Italy, Alliance of Liberals and Democrats for Europe

Summary

Social media allow individual users to share content with a large audience and to engage in virtual communities and networks. These activities are contractually regulated by the content moderation rules contained in social media companies' terms and conditions, and users are bound by them on a take-it-or-leave-it basis.

Member States must establish the basic principles and institutional framework that can correct the power imbalance resulting from the unequal contractual relationship and ensure the effective protection of the right to freedom of expression online. It is imperative, however, that public regulation of content moderation does not lead to an overzealous approach to content removal by social media companies.

The report calls on social media companies to refrain from implementing policies that unduly restrict users' freedom of expression. Their terms and conditions must be clear, easily accessible and based on fundamental rights principles. They must provide human moderators with comprehensive training and adequate working conditions (including mental health care), and make effective use of automated content moderation tools that are subject to human oversight and rigorous and ongoing evaluation. They must inform users promptly and in a reasoned manner of any content moderation action taken. Their complaint handling systems must be easily accessible and user-friendly, and they must support the establishment of independent out-of-court dispute resolution bodies.

[1] Reference to committee: Doc. 15555, Reference 4705 of 23 January 2023.

A. Draft resolution[2]

1. Social media have become an online agora where users come to exercise their right to freedom of expression and information in many ways. These include posting their own content and enjoying the content posted by others, getting informed and informing others, and communicating with other users.

2. The right to freedom of expression is not an absolute right; social media are legally obliged to remove any illegal content when they become or are made aware of its existence on their services. Moreover, it is incumbent upon social media to combat the dissemination of harmful content.

3. Social media companies are also bearers of fundamental rights, such as rights of property and freedom of enterprise, and therefore they have a say on how users can use their services and on what content they can post. The content moderation rules included in their terms and conditions (T&Cs) allow for social media to demote, demonetise, restrict access to, or remove a concrete content item because of its incompatibility with their T&Cs. In extreme cases, social media companies can suspend or even terminate a user's account. Their T&Cs have a contractual character, and users are bound by them on a take-it-or-leave-it basis.

4. The major social media companies, mainly US-owned, have a global reach; their content moderation policies and their commercial or ideological decisions about content to promote or demote may have an immense influence on public opinion and on choices of billions of people. It is, nevertheless, incumbent upon them to respect the laws of the country in which they provide their services.

5. Given the potential impact on societal behaviours and on the proper functioning of democratic processes that the information and communication flow on social media de facto has, it is incumbent upon the state to establish the fundamental principles and institutional framework that may correct the power imbalance resulting from the unequal contractual relationship and ensure the effective protection of the right to freedom of expression.

6. It is imperative, however, that public regulation of content moderation does not have a chilling effect on free speech and is not intended to impose the views of the political power in place and a censorship on opinions or ideas which may conflict with the ruling majority's vested interests. Moreover, national regulations should not place undue burdens on social media, which could result in an overzealous approach to content removal. These regulations and their implementation must uphold freedom of expression and carefully assess the necessity of any restrictions.

7. The risk of restrictive content moderation policies is increased by the lack of transparency in their implementation. Social media have been accused of a practice called "shadow banning" whereby they delist or demote content dealing with controversial issues without notifying the user in question, making that content invisible to other users. This devious, hidden practice should be forbidden: it deprives users of the possibility to defend effectively their right to freedom of expression.

8. The press and the media in general use social media as a platform for disseminating information to the public. It is therefore essential that content moderation practices do not unduly impact media and journalistic content that respect professional standards and the national regulatory framework.

9. Content moderation is increasingly carried out by automated means. Artificial intelligence (AI) tools are much more efficient than human moderators in processing at a high speed the colossal amount of content circulating on the web, to identify prohibited content. They lack to date, however, the capacity to fully understand the subtleties of human interaction (humour, parody, satire, etc.) and to assess the content in its context.

10. For this reason, human moderators must remain the cornerstone of any content moderation system and be responsible for making decisions in cases where automated systems are not up to the task. However, human moderation can be biased and lead to inconsistencies among countries due to cultural differences; it is therefore imperative to establish clear and comprehensive standards and to guarantee appropriate training, to ensure that all moderators have the requisite knowledge of both the applicable legislation and the company's internal guidelines.

---

[2] Draft resolution adopted unanimously by the committee on 4 December 2024.

11. Regrettably, despite their fundamental role, human moderators' working conditions are inadequate, they are overexposed to disturbing content that can cause them serious mental health problems and they suffer from restrictions on their freedom to speak out about the problems they encounter at work.

12. Generative artificial intelligence (GenAI) tools allow to produce synthetic content that is virtually indistinguishable from human generated content. Such content can be highly misleading, be a tool of disinformation and manipulation, and instigate hatred and discrimination, among other dangers. It is essential that users are made aware of content that appears to be genuine, but which is in fact not. In this regard, watermarking techniques are particularly beneficial but have several drawbacks, including their lack of interoperability among social media services.

13. Independent assessment of T&Cs and content moderation policies and their enforcement, also with a view to identifying and promoting best practices, could help to ensure their consistency with principles which uphold a human-rights approach to content moderation.

14. The establishment of clear and transparent rules for conflict resolution is essential to ensure the protection of users and to minimise the risk of being subjected to a potentially biased decision by the social media company, or of being forced to pursue costly legal action against a multinational corporation with enormous financial resources at its disposal.

15. The establishment of independent out-of-court dispute settlement bodies to assess content moderation decisions may prove beneficial in enhancing compliance with fundamental rights. Collaboration between social media companies in establishing such bodies could also hopefully facilitate dispute resolution.

16. As recalled by the Assembly in Resolution 2281 (2019) "Social media: social threads or threats to human rights?", social media companies should employ algorithms that promote diversity in sources, topics and views, guarantee the quality of information available, and thereby reduce the risk of filter bubbles and echo chambers.

17. In light of these considerations, the Assembly calls on member States to review their legislation to better safeguard the right to freedom of expression on social media. In this respect, they should in particular:

17.1. require that social media uphold users' fundamental rights, including freedom of expression, in their content moderation policy and implementation practices;

17.2. require that social media platforms provide justification for any measure taken to moderate content provided by the press or media service providers prior to its implementation and allow them an opportunity to reply within an appropriate timeframe.

17.3. provide for minimum standards of working conditions for human moderators, including a requirement of adequate training to carry out their often stressful tasks and of access to proper psychological support and mental health care when needed;

17.4. sign and ratify the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law (CETS No. 225, "the Vilnius Convention") and adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place to meet the challenges of the identification of content generated by artificial intelligence systems;

17.5. require that AI-generated content is disclosed as such by those initially posting it and that social media implement technical solutions allowing for such content to be easily identified by users, and encourage collaboration between social media companies to ensure the interoperability of watermarking techniques for AI-generated content;

17.6. require that out-of-court dispute settlement bodies, when established, are independent and impartial, have the necessary expertise, are easily accessible, and operate according to clear and fair rules, with certification of these requirements by the competent national regulatory authority.

17.7. promote, within the Internet Governance Forum and the European Dialogue on Internet Governance, reflection on the possibility for the internet community to develop, through a collaborative and, where appropriate, multi-stakeholder process, an external evaluation and auditing system aimed at determining whether algorithms are unbiased and respect the right to freedom of expression, and a

"seal of good practices" which could be awarded to social media whose algorithms are designed to reduce the risk of filter bubbles and echo chambers and to foster an ideologically cross-cutting, while safe, user experience.

18. The Assembly calls on social media companies to avoid measures that unnecessarily restrict the freedom of expression of users. They should, in particular:

18.1. directly incorporate principles of fundamental rights law, and in particular freedom of expression, into their T&Cs;

18.2. use caution when moderating content that is not obviously illegal;

18.3. provide users with T&Cs that are readily accessible, clear and informative on the types of content that are permissible on their services and the consequences for non-compliance, and which are understandable to the wide span of users notwithstanding differing levels of digital literacy and reading proficiency;

18.4. notify users without undue delay of any moderation action taken on their content, providing a comprehensive account of the rationale behind the decision, accompanied by a reference to the internal rules which have been applied;

18.5. refrain from shadow banning users' content and notify users of every instance of demotion or delisting;

18.6. ensure that automated content moderation processes are subject to human oversight and to rigorous and continuous evaluation to assess their performance;

18.7. make available a system for handling complaints that is easily accessible, user-friendly, and allows users to make a precise complaint;

18.8. give human moderators appropriate training and working conditions which pay attention to the heavy psychological stress they are submitted to, and ensure adequate protection to their health;

18.9. refrain from permanent deletion of content (including its metadata) that has been removed in accordance with legal obligations or with T&Cs, in particular when the content in question may serve as evidence of war or other crimes;

18.10. ensure that the AI systems they develop or use uphold Council of Europe standards, including the new Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law; algorithms should be designed to respect the right to freedom of expression, and to encourage plurality and diversity of views and opinions while ensuring a safe user experience, their operation modalities should be disclosed and, users duly informed on how these algorithms filter and promote content;

18.11. collaborate with other online services with the aim of ensuring the interoperability of watermarking techniques for AI-generated content;

18.12. promote and support the creation of independent out-of-court dispute settlement bodies, and abide by their decisions and recommendations;

18.13. support the work of independent third-party oversight bodies and abide by their decisions and recommendations;

18.14. ensure that decisions related to content moderation are duly motivated and that researchers have access to full information on the legal base and reasoning behind each decision.

B. Explanatory memorandum by Ms Valentina Grippo, rapporteur

1. Introduction

1. Social media[3] have the power - and in some cases the legal obligation - to filter, prevent the dissemination of, downgrade, block and, where appropriate, take down illegal or potentially harmful content. In addition, their terms and conditions (T&Cs) set out rules about what content and behaviour is acceptable on their services and how they can restrict the provision of those services. However, certain implementation measures taken as part of their content moderation policies may conflict with the right to freedom of expression.

2. While social media companies are themselves bearers of fundamental rights such as right to property and freedom of enterprise, it falls to public regulation to lay down the fundamental principles and institutional framework needed to ensure the effective protection of users' fundamental rights, in particular the right to freedom of expression.

3. The objective of my report is to examine how the right to freedom of expression on social media can be better protected, while ensuring that the rights to property and freedom of enterprise of social media companies are not unduly infringed upon.

4. To this end, I will consider the following issues:

- The obligations of social media companies under EU law, regarding the content, application and enforcement of their T&Cs, and in particular their content moderation measures and the mechanisms for checking and ensuring compliance with those obligations;
- Specific measures advocated under the Council of Europe's soft law;
- The content moderation policies of major social media companies and the possible difficulties which they encounter in taking EU legislation and Council of Europe standards onboard;
- Standards and measures which should be introduced to take account of recent and upcoming developments.

5. My analysis builds on contributions from several experts we heard from,[4] and benefits from the insights gathered during my July 2024 fact-finding visit to Dublin, where I met with representatives of three major tech companies: Meta, TikTok and Google. I also had fruitful meetings with the Irish media and data protection regulators.

2. The regulation of content moderation under EU law

6. Since the year 2000, online services have been regulated at the EU level by the Directive on electronic commerce. This Directive includes an exemption of liability for hosting providers when they do not have actual knowledge of illegal content or activity or when they act promptly once they gain such knowledge. Member States, however, cannot impose a general obligation on information society service providers to monitor the information which they transmit or store, nor a general obligation to actively seek facts or circumstances indicating illegal activity.

7. These general rules have been complemented over the period 2000-2020 by specific EU legislation and co-regulation covering audiovisual media services, copyright, terrorism, child sexual abuse, hate speech, and disinformation.

8. In 2019, the European Commission launched the process for the adoption of a comprehensive regulatory package, the "Digital Services Act package". As a result of this process, two new Regulations, the Digital Services Act and the Digital Markets Act, were enacted in 2022.

---

[3] For the purposes of this report, the term 'social media' refers to online services that enable individual users to share content with a vast audience and engage in virtual communities and networks. This definition encompasses platforms such as Facebook, Tik Tok, YouTube, and similar services. Despite this working definition, the terms 'platforms' and 'online platforms' are used throughout the text to align with the terminology used in other contexts, particularly in EU legislation.

[4] I wish to thank in particular: Mr Lubos Kuklis, Digital Services Act (DSA) team, European Commission, Brussels; Mr Jack Goodman, BBC, United Kingdom; Mr Christian Hannibal, Head of Public Policy, Tik Tok, Denmark; Mr Mark David Cole, Professor for Media and Telecommunication Law at the University of Luxembourg; and Ms Gemma Shields, Online Safety Policy Lead (Human Rights and Transparency) at Ofcom, United Kingdom.

2.1. The Digital Services Act

9. The Digital Services Act (DSA) aims to prevent illegal and harmful activities online and the spread of disinformation on intermediary services. The DSA has a special focus on so-called "Very large online platforms" (VLOPs) and "very large online search engines" (VLOSEs), which, due to their size in terms of recipients of these services,[5] are subject to special regulation.

10. Article 14 DSA regulates the activities of intermediary services (including social media) aimed at detecting, identifying and addressing content provided by users that is incompatible with the provider's T&Cs.

11. As a general principle, the DSA upholds the freedom of contract of providers of intermediary services, but sets rules on the content, application and enforcement of their T&Cs "in the interests of transparency, the protection of recipients of the service and the avoidance of unfair or arbitrary outcomes" (Recital 45 DSA). Their T&Cs must clearly indicate the grounds on which they may restrict the provision of their services to users.

12. Intermediary services must explain in their T&Cs the restrictions they impose concerning content published by their users and how they moderate that content. They must in particular include information on any policies, procedures, measures, and tools used for the purpose of content moderation, including algorithmic decision-making and human review, as well as the rules of procedure of their internal complaint-handling system. This information must be provided in clear, plain, intelligible, user-friendly, and unambiguous language, and must be publicly available in an easily accessible and machine-readable format.

13. Intermediary services must apply and enforce these rules in a diligent, objective and proportionate manner, with due regard to the rights and legitimate interests of all parties involved, including users' fundamental rights such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter of Fundamental Rights of the EU.

14. Furthermore, intermediary services must inform users of any significant change to the T&Cs.

15. Intermediary services primarily directed at or predominantly used by minors must have T&Cs that minors can understand.

16. With regard to recommender systems, T&Cs must set out in plain and intelligible language their main parameters, including how the information suggested to the user is determined, and the reasons for the relative importance of those parameters. T&Cs must also describe the options available for users to modify or influence those main parameters. Where several options are available to determine the relative order of information presented to users, the service must also allow the user to select and to modify at any time their preferred option in an easy way (Article 27 DSA).

17. Content moderation decisions must be notified to users with a statement of reasons (Article 17 DSA). Moreover, effective complaint-handling systems must be in place (Article 20 DSA), and users must be entitled to select any out-of-court dispute settlement body that has been certified by the relevant Digital Services Coordinator (Article 21 DSA).

18. EU Member States should respect the fundamental rights to an effective judicial remedy and to a fair trial as provided for in Article 47 of the EU Charter of Fundamental Rights and in Articles 6 and 13 of the European Convention on Human Rights (ECHR).

19. VLOPS and VLOSEs must assess the systemic risks stemming from the design, functioning and use of their services, as well as from potential misuses by the users, and should take appropriate mitigating measures in observance of fundamental rights (Article 34 DSA).

20. When conducting this risk assessment, VLOPs and VLOSEs must take into account amongst other things, the impact of their content moderation systems and the applicable T&Cs and their enforcement.

---

[5] VLOPs and VLOSEs are online platforms and online search engines which have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million, and which are designated as such by the European Commission (Article 33(1) DSA).

21. VLOPS and VLOSEs must put in place reasonable, proportionate, and effective mitigation measures, tailored to the specific systemic risks mentioned above, with particular consideration given to the impacts of such measures on fundamental rights. Such measures may, among others, include adapting their T&Cs and their enforcement and content moderation processes (Article 35 DSA).

22. Intermediary services must ensure an adequate level of transparency and accountability (Recital 49 DSA). To this effect, they must make publicly available clear, easily comprehensible reports on their content moderation activities. These reports must be published at least once a year, in a machine-readable format and in an easily accessible manner (Article 15 DSA). More stringent reporting obligations apply to providers of online platforms (Article 24 DSA) and especially to VLOPs and VLOSEs (Article 42 DSA), which are also subject to independent audits (Article 37 DSA).

23. The European Commission maintains the DSA Transparency Database which contains the decisions and statements of reasons of online platforms when they remove or otherwise restrict availability of and access to information. Its aim is to ensure transparency and to enable scrutiny over the content moderation decisions of online platforms and to monitor the spread of illegal content online (Recital 66 DSA).

2.2. The European Media Freedom Act

24. The European Media Freedom Act (EMFA) regulates media pluralism and independence in the EU.

25. Regarding content moderation, Article 18 EMFA protects media service providers (e.g. TV broadcasters) against the unjustified removal by VLOPs of media content considered incompatible with their T&Cs. VLOPs will need to explain the reasons for the content moderation measure before it takes effect, and should give the media service provider an opportunity to reply within 24 hours, or within a shorter timeframe in crisis situations referred to in the DSA. This early warning procedure does not apply, however, when content moderation decisions are taken following the rules of the DSA and the AVMSD, notably the obligations to remove illegal content, protect minors, and mitigate systemic risks.

3. Standard-setting of the Council of Europe

26. The Parliamentary Assembly and our committee in particular have been addressing the issue of regulation of freedom of expression on the Internet for several years. As far back as in 2012, in our report on "The protection of freedom of expression and information on the Internet and online media", we voiced fears that private operators with dominant positions on the internet services market might unduly restrict access to, and dissemination of, information without informing their users and in breach of user rights.

27. In Resolution 1877 (2012), the Parliamentary Assembly therefore called on member States to ensure that internet intermediaries be transparent (§ 11.3.) and held accountable for violations of their users' right to freedom of expression and information (§ 11.6.).

28. Subsequently, other reports by our committee considered the issue of public regulation of freedom of expression on the Internet from various angles and the Parliamentary Assembly developed guidelines concerning both the drafting of national legislation and private operators' self-regulation standards (see AS/Cult/Inf (2024) 12).

29. Content moderation and the impact of the algorithms used by internet platforms on human rights have also been covered in detailed studies by the Council of Europe intergovernmental sector. Standards aimed at clarifying the responsibilities of public authorities and also at regulating the action of Internet intermediaries are set out in Recommendation CM/Rec(2018)2 on the roles and responsibilities of Internet intermediaries, CM/Rec(2022)13 on the impact of digital technologies on freedom of expression, CM/Rec(2022)16 on combating hate speech, and Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems. Moreover, the Guidance Note on content moderation (2021) and the Guidance Note on the prioritisation of public interest content (2021) include principles and lines of action which uphold public access to quality information. Also, the Guidance Note on countering the spread of online mis- and disinformation through fact-checking and platform design solutions in a human rights compliant manner (2023) focuses on three areas of action: fact-checking, platform design solutions, user empowerment, and media literacy.

30. Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries provides that "States have the ultimate obligation to protect human rights and fundamental freedoms in the digital environment. All regulatory frameworks, including self- or co-

regulatory approaches, should include effective oversight mechanisms to comply with that obligation and be accompanied by appropriate redress opportunities" (guideline 1.1.3.).

31. The most detailed standard-setting document of the Council of Europe is its Guidance Note on Content Moderation (2021). It provides practical guidance to member States of the Council of Europe, taking into account existing good practices for policy development, regulation and use of content moderation in the online environment in line with their obligations under the ECHR. The Guidance note is also addressed to internet intermediaries who have human rights responsibilities of their own.

32. The Guidance Note lists key principles for a human-rights-based approach to content moderation:

a. Transparency – Transparency is essential for ensuring accountability, flexibility, non-discrimination, effectiveness and proportionality, as well as for the identification and mitigation of conflicts of interest. Minimum standards should be identified to assess whether the content moderation in question is achieving its specific goals, and there should be an independent review of at least a representative sample of content moderation cases.

b. Human rights by default – Under the ECHR, human rights are the rule and restrictions are the exception, and this approach must guide the development of policies in relation to content moderation. It is necessary to proactively identify which rights might be threatened before content is moderated. Also, the right to effective remedy (Article 13 ECHR) must be upheld. Moreover, prior review of self- or co-regulatory measures is not enough to ensure that human rights are respected, so frequent review of the impact(s) of measures is essential.

c. Problem identification and targets – Policy interventions that have the purpose of minimising risk must be clearly recognised as such, in order to mitigate the particular problems of this approach, with the state taking its share of responsibility. These policy interventions should have clear targets, adjustment mechanisms and supervision, meaningful protection for freedom of expression, as well as tools to identify counterproductive impacts. If the content moderation is being carried out in the context of a self- or co-regulatory scheme, there should also be mechanisms built in, to redesign, adapt or abandon the scheme, when minimum standards are not achieved or if the nature of the problem evolves in a way which makes the identified approach not effective.

d. Meaningful decentralisation – Decentralised, multi-stakeholder, remunerated, empowered and independent moderation is essential to take regional peculiarities into account when dealing with the most difficult types of content. Online platforms should develop multi-stakeholder councils to help them evaluate the hardest problems, to evaluate emerging issues, and to dissent to the highest levels of company leadership. Furthermore, adequate data needs to be made available to civil society and technical and academic researchers to facilitate ongoing analysis.

e. Communication with the user – Restrictions to human rights should respect human rights norms and be as transparent as possible. T&Cs should be as clear and accessible as possible. The application of those rules should also be predictable, in line with human rights law. Also, users must be able to file a complaint with the company in the most specific possible way. If content removal is not urgent, the users concerned should be given clear information about why their content may have breached T&Cs or the law, should have the right to defend their case within a set timeframe and, in any case, the right to a meaningful appeal. Content that needs to be taken offline as quickly as possible must be well defined and the process for reviewing it, deleting it and, when necessary, putting it back online, needs to be predictable, accountable, and proportionate.

f. High level administrative safeguards – A clear and predictable legal framework is essential to ensure that restrictions are provided for by law. This also requires a meaningful transparency on governance, decision-making processes, and details of how, when, why and how often, what content was removed, or not, and for what reason. Good transparency reporting will allow both companies and the public to be able to identify the gaming[6] of companies' complaints mechanisms. Also, the consistency and independence of review mechanisms requires public availability of enough data on decisions and enough sample cases made available to an independent and impartial body for proactive review. The

[6] This is the "deliberate, not necessarily illegal, manipulation or abuse of internet intermediaries' complaints systems". See page 25 of the Guidance Note.

findings must be meaningfully considered by the internet intermediaries. Furthermore, due attention must also be given to the labour rights and mental health of all workers involved in moderation of content which may be shocking, disturbing, or otherwise likely to have a psychological impact on them. Privacy and data protection must be ensured. States should consider the rights of victims of illegal content to ensure full support to negate or mitigate the damage caused. Appropriate measures are also needed to compensate victims of unjustified takedowns and to avoid such problems arising.

g. Addressing the peculiarities of self- and co-regulation in relation to content moderation online – The scale of the communication flow with users differs significantly when looking at online media. As a result, assumptions based on experience of traditional media self- and co-regulation may be misleading in the context of most internet intermediary self-regulation. It is crucial that the role of the State be properly defined, to ensure accountability.

4. Content moderation in practice

4.1. Practices in comparison

33. Content moderation is a term that englobes all activities undertaken by intermediary services with the aim to detect, identify and address user-generated content that is either illegal or incompatible with their T&Cs. In practice, this aim is enforced by measures that affect content (demotion, demonetisation, disabling of access, removal), and measures that affect the user (termination or suspension of a user's account).[7]

34. The content moderation policies, rules for enforcement, and enforcement practices of each social media are specific to each of them, which makes it challenging to conduct a comparative analysis. Furthermore, it is also important to exercise caution when examining the transparency reports of these social media as there are several methodological differences, the observation periods do not always correspond to the six months required by the EU, and a significant amount of information is missing.[8]

35. A 2024 study on media and society after technological disruption reviews the content-moderation policies and enforcement practices of Facebook, YouTube, TikTok, Reddit, and Zoom. It discusses content policies, enforcement rules, and enforcement implementation and transparency.

36. The three largest social media (Facebook, YouTube, and TikTok), which face similar challenges given their size, have similar policies on content moderation:

- Platform-wide policies against given types of content;
- Tiered approaches to enforcement, involving banning some kinds of content and limiting access to or distribution of other kinds of content;
- A policy of warning users who post violative content and banning those users who do so repeatedly.

37. Reddit's channel-specific approach is very different, as most policy rules are set by users themselves according to the types of discussions they want to have within specific groups.

38. Social media usually enforce their content-moderation policies proactively by looking for content that violates policies, but some do so reactively in response to user reports.

39. Regarding the consequences on users violating T&Cs, most social media employ warning systems of some kind, but with not enough clarity for the user, YouTube being a positive notable exception in this respect. The lack of clarity may allow social media to adjust their policies in reaction to events without having to communicate every change publicly, but this reduces transparency and accountability.

40. Similarities in policy enforcement appear around illegal content since legal requirements set the framework, although the enforcement mechanisms may differ. Each of the social media considered here makes clear in their T&Cs that users posting illegal content will lose their accounts immediately, without strikes or warnings.

---

[7] See Article 3(t) DSA.
[8] See Content moderation: Key facts to learn from Facebook, Instagram, X and TikTok transparency reports.

4.2. Concerns

41. Content moderation implies a clash of rights between users' fundamental rights, such as freedom of expression, and social media companies' rights to property and freedom of enterprise. In principle, social media have a right to decide what content is acceptable on their services and how users can use their services. These rules are contained in their T&Cs, which have a contractual character, and users are bound by them on a take-it-or-leave-it basis, without the possibility of negotiating any clause.

42. This regulation of content by contract raises serious concerns in terms of protection of freedom of expression for the following reasons:[9]

- Lower free speech standards: While social media are in principle free to restrict content based on freedom of contract, certain policies, such as the removal of content 'for any reason or no reason', fall short of the responsibility to respect human rights. It is also problematic that the rules governing content moderation are not guided by the principles of necessity and proportionality.
- Lack of transparency and accountability: Lack of transparency regarding the implementation of T&Cs impacts the ability to hold companies accountable for wrongful, arbitrary, or discriminatory content takedowns. This may be due for example to lack of clarity about the scope of a content-related rule, lack of availability of T&Cs in certain languages, or lack of granularity in content moderation reports.
- Lack of procedural safeguards and remedy: Examples include lack of clarity concerning when users are notified that their content has been moderated, or whether their account has been penalised in any way, and the reasons for such actions; notifications that are often too general, simply referring to a policy that has allegedly been breached without any further justification, or simply not made at all;[10] lack of adequate legal remedies in T&Cs, with problematic dispute resolution and choice of law clauses that may deter most users from bringing litigation for example in their local courts.
- Circumventing the rule of law: Sometimes, authorities contact social media companies informally to request the removal of content on the basis of the companies' T&Cs. While companies may not be legally obliged to comply with such requests, they may decide on such request to remove the content without giving users the opportunity to challenge the legality of the restriction in question in court.

43. In this regard, the press and news media outlets require specific regulatory measures. These services play a pivotal role in the exercise of the right to receive and impart information online, and they operate in accordance with relevant legislation and professional standards. In the case of broadcast media, they are also subject to national regulatory oversight. It thus follows that when a social media provider moderates content from press or news media outlets, special consideration should be given to the principles of media freedom and media pluralism.[11] Following the EMFA as a model, social media providers should provide an explanation of the rationale behind the moderation measure in question prior to its implementation. Furthermore, they should afford the press or media outlet in question the opportunity to provide a right of reply.

44. Of particular concern is the moderation of videos shot in war zones. Given their violent and disturbing nature, social media companies usually remove these videos to protect their users. Beyond their informational value, however, these videos could also serve as evidence of war crimes in a court of law. This issue was already raised during the Syrian war in 2011, and has come to the fore in particular during the Russian war of aggression against Ukraine.[12] The use of automated content moderation tools seems to have exacerbated the problem, as they lack the capacity to interpret the value of this war footage.[13] While there are private organisations such as The Syrian Archive and Mnemonic that collect, archive, and research this kind of footage, social media companies should do more in order to avoid removing war footage unless absolutely necessary, and to preserve it (including its metadata) when the content in question could serve as evidence of war crimes. The problem is, however, wider, and social media companies should seek to avoid permanent deletion of evidence which could be used to pursue and sanction other crimes too.

---

[9] See Content moderation and freedom of expression handbook.
[10] For more information on the practice of "shadow banning" see below.
[11] For cases where social media failed to give appropriate consideration to media freedom and media pluralism in their content moderation activities see Is big tech tampering with media content!?.
[12] See The difficulty of conserving social media evidence of war crimes.
[13] See AI: War crimes evidence erased by social media platforms.

5. Options for conflict resolution

45. Prior to the introduction of the DSA, users in the EU had just two options if they wanted to challenge content moderation measures: either file a complaint with the service's internal complaint-handling system (if this option was available), with a risk of a potentially biased decision by the service or go down the judicial path pursuing costly legal action.[14]

46. The DSA has established a comprehensive regulatory framework governing three conflict-resolution avenues: lodging a complaint before an internal complaint-handling system, applying to an out-of-court dispute settlement body, and bringing the case to justice.

5.1. Internal complaint-handling systems

47. Based on Article 20 DSA, online platforms must provide users with a system to lodge complaints, electronically and free of charge, against content moderation decisions that negatively affect them. Users have six months to apply to the system since the day they were informed of the platform's decision. The internal complaint-handling system must be easily accessible, user-friendly, and enable and facilitate the submission of sufficiently precise and adequately substantiated complaints. Online platforms must handle complaints in a timely, non-discriminatory, diligent and non-arbitrary manner, and reverse unsubstantiated decisions without undue delay. Users must be informed without undue delay of the reasoned decision and of the possibility of out-of-court dispute settlement and other available possibilities for redress. The decisions must be taken under the supervision of appropriately qualified staff, and not solely on the basis of automated means.

5.2. Out-of-court dispute settlement bodies

48. The real novelty, introduced by the Article 21 DSA, is the possibility for users to select a certified out-of-court dispute settlement (ODS) body, to resolve disputes relating to content moderation decisions. However, the decisions made by certified ODS bodies are not binding.

49. Information about the possibility to have access to an ODS body must be easily accessible on the platform's online interface, clear and user-friendly.

50. The recourse to an ODS body is without prejudice to the user's right to initiate, at any stage, judicial proceedings before a court of law. In any event, online platforms may refuse to engage with an ODS body if a dispute has already been resolved concerning the same information and the same grounds of alleged illegality or incompatibility of content.

51. ODS bodies are certified by the Digital Services Coordinator[15] (DSC) of the EU Member State where they are established. They must be impartial and independent, have the necessary expertise, be remunerated in a way that is not linked to the outcome of the procedure, be easily accessible online, be capable of settling disputes in a swift, efficient, and cost-effective manner, and their procedural rules must be clear, fair, easily and publicly accessible, and comply with applicable law. ODS bodies must report annually to the DSC on their functioning. DSCs shall, every two years, draw up a report on the functioning of the ODS bodies that they have certified.

52. Certified ODS bodies must make their decisions available to the parties within a reasonable period of time and no later than 90 calendar days after the receipt of the complaint. In the case of highly complex disputes, this period up may be extended to 180 days.

53. Regarding procedural costs, the platform must cover them irrespective of the outcome of the decision as long as the user has engaged in good faith. This asymmetrical system could have the beneficial effect of ensuring that online platforms will tend to minimise their ODS procedural costs by making fewer errors in moderating content.[16]

---

[14] See Resolving content disputes outside the courtroom using the Digital Services Act.
[15] Digital Services Coordinators are national authorities responsible for all matters relating to supervision and enforcement of the DSA in a EU member State, see Article 49 DSA.
[16] See Settling DSA-related Disputes Outside the Courtroom: The Opportunities and Challenges Presented by Article 21 of the Digital Services Act (quoting Principles of the Digital Services Act).

54. Meta did not wait for the adoption of the DSA in order to introduce an independent body for conflict resolution called the Oversight Board (OB). OB is composed of independent members who are not Meta employees and cannot be removed by the company. Meta does not play a role in the selection of new OB Members.

55. When Facebook, Instagram or Threads users have exhausted Meta's appeals process, they can challenge the company's moderation decision by appealing to the OB. Meta can also refer cases to the OB. The OB can choose to overturn or uphold Meta's decisions, and Meta is committed to implementing the Board's decisions.

56. Furthermore, the OB makes non-binding recommendations on how Meta can improve its content policies, enforcement systems and overall transparency for Facebook, Instagram, and Threads. These recommendations are public, and the OB's Implementation Committee, which is made up of Board Members, uses both publicly available and internal data to understand the impact of their recommendations.

57. The OB publishes transparency reports which include details about the impact of the recommendations on users, the decisions taken and the cases that the OB is receiving from users. The OB also publishes annual reports that assess Meta's performance in implementing OB's decisions and recommendations. The most essential tool for gathering the metrics included in each report is the Recommendation Tracker.

58. A more recent example shows how collaboration between social media companies will hopefully facilitate dispute resolution. Appeals Centre Europe is a new, certified ODS body[17] that was launched in late 2024 and at the start will settle disputes relating to Facebook, TikTok and YouTube, making decisions on whether the company's decision is consistent with its own content policies, including any rules or exceptions that refer to human rights. The Appeals Centre Europe will operate with an in-house team of experts to resolve disputes, applying human review to every case. Complex cases will be reviewed by specialists with expertise in specific regions, languages, or policy areas.

59. The Appeals Centre Europe has been set up through a one-time grant from Meta's Oversight Board Trust and will be funded through fees. Social media companies that will participate in the Appeals Centre will pay a fee in connection with every case, while users who appeal will only pay a nominal fee, which will be refunded when the Appeals Centre's decision is in their favour.[18]

5.3. Judicial redress

60. The fundamental right to an effective judicial remedy and to a fair trial, as provided for in Article 47 of the EU Charter of Fundamental Rights and Articles 6 and 13 ECHR, apply to cases where there are disputes around the content moderation activities of social media companies. The DSA does not prevent relevant national judicial or administrative authorities from issuing an order to restore content where such content complied with the platform's T&Cs but had been erroneously removed (Recital 39 DSA).

6. Finding the right balance of rights

61. As stated in the Council of Europe's Guidance Note, content moderation entails the imposition of limitations on fundamental freedoms, and it is incumbent upon states to guarantee that the restrictions imposed by regulatory, self- and co-regulatory regimes are transparent, justified, essential and proportionate. In the case of T&Cs, the difficulty lies in distinguishing between legitimate co-regulation, illegitimate state pressure on private companies and legitimate enforcement by social media of their internal rules.

62. As recognised by the UN Framework and Guiding Principles on Business and Human Rights, companies have a corporate responsibility to respect human rights, that is, acting with due diligence to avoid infringing the rights of others and addressing harms that do occur. A company's responsibility to respect human rights applies across all its business activities and relationships.

63. Moreover, as explained in Recital 45 DSA, while the freedom of contract of providers of intermediary services should in principle be respected, it is nonetheless not absolute and must respect certain rules on the

[17] Legally certified under Article 21 DSA by the Irish media regulator Coimisiún na Meán.
[18] For further information see EU creates 'Appeals Centre' to referee disputes with social media giants.

content, application, and enforcement of the T&Cs of those providers in the interests of transparency, the protection of recipients of the service and the avoidance of unfair or arbitrary outcomes.

64. In Europe, the number of court cases dealing with conflicts regarding content moderation is on the rise, offering insights into significant legal issues. In general, courts have opted not to directly apply the right to freedom of expression as a basis for their decisions. Instead, they have resorted to applying fundamental legal principles, such as good faith, fairness, breach of contract, consumer protection, and even data protection.

65. On 29 July 2021, the German Federal Court of Justice issued an important judgment in which it discussed the balancing of rights between social media and users.[19] The court explained that Facebook, as a private company, is not directly bound by fundamental rights, and therefore not bound by the right to freedom of expression (Article 5 (1)(1) of the German Basic Law) in the same way as the State. Pursuant to Article 1(3) of the Basic Law, fundamental rights only bind the legislature, the executive, and the judiciary as directly applicable law. Moreover, Facebook's dominant position in the field of social media does not result in a state-like obligation. In particular, Facebook does not assume the provision of framework services for public communication, which the Federal Constitutional Court considers a prerequisite for a private entity to be bound by fundamental rights as the State. In addition, Facebook' itself is a bearer of fundamental rights: in civil disputes between private individuals, their fundamental rights must be understood in their interaction and balanced in such a way that they are as effective as possible for all parties involved.

66. A similar interpretation was made by a Dutch court in a judgment of 6 October 2021.[20] In this case, which concerned posts containing Covid misinformation, the court explained that the European Court of Human Rights (ECtHR) does not, in principle, grant direct horizontal effect to the provisions of Article 10 ECHR. Even if it had been the case, the court would have had to consider that, in the case at hand, this fundamental right competed with LinkedIn's fundamental rights. The court ruled that LinkedIn's policy regarding Covid-19 was vague, and the company should restore the user's account since it had failed to provide safeguards regarding the termination of users' accounts. However, the misinformation messages in question were not required to be re-uploaded since LinkedIn had compelling reasons to conclude that they contained harmful misinformation.

67. There are further cases in which courts ruled against the termination or suspension of a user's account on the basis of the vagueness or unfairness of the service's contractual clauses. For example, on 5 June 2024, a French court ruled that Facebook's closing of a historian's account was made in breach of contract and that the rule applied by the social network was unfair.[21] In the case at hand, the historian had published on its Facebook page an article concerning the activities of Daesh in Algeria. Following this publication, Facebook deactivated the historian's account providing as reason only a generic email where they recalled their policy of not allowing credible threats to harm others, support for violent organisations or excessively outrageous publications. The court, however, concluded that the historian had unambiguously denounced the terrorist group, and that the mere reproduction of a Daesh press release could not be considered as an endorsement of their actions in view of the contextualisation made within the publication. Therefore, this publication did not fall within the scope of the unauthorised actions on the social network and could not be deemed to fulfil the conditions set out in Facebook's terms of service for suspending or terminating an account. Moreover, the court ruled that said clause was abusive according to Article R.212-2 of the French Consumer Code, which states that "in contracts concluded between professionals and consumers, clauses whose object or effect is to (...) 4° Giving the trader the right to terminate the contract without reasonable notice are presumed to be unfair, unless the professional can prove otherwise".[22]

68. It is important to recall that content moderation is a challenging process, primarily due to the vast quantity of content that requires moderation and the intricate legal assessment that must be conducted on a case-by-case basis in numerous instances. Also, the increasing social pressure to fight harmful content plus commercial interests may lead to over-reaction from the social media's side.[23] Anyone whose content is

---

[19] See BGH, III ZR 179/20, 29. Juli 2021 and BGH III ZR 192/20. For an in-depth description (in English) of the case see The Case on Facebook's Terms of Service.
[20] See Rechtbank Noord-Holland, C/15/319230 / KG ZA 21-432, 06-10-2021. See also Van Haga v. LinkedIn.
[21] Tribunal judiciaire de Paris - 17ème Ch. Presse-civile 5 juin 2024 / n° 21/00726. See also [FR] Meta breached contract by closing Facebook account of historian who denounced Daesh abuses.
[22] For a similar case regarding Covid-19 misinformation see Van Haga v. YouTube.
[23] See e.g. Remedying Overremoval.

considered shocking, controversial, or otherwise undesirable (even if legal) can be affected by such overreactions.

69. In this regard, a particularly problematic type of content moderation measure is the so-called "shadow banning". This is a practice whereby social media delist or demote content without notifying the user in question. The content remains accessible, yet the user is unaware that it is essentially invisible to other users and therefore not accessed.

70. It appears that shadow banning is employed by social media companies primarily in the context of controversial matters. For example, since the Hamas attack on Israel on 7 October 2023 and the resulting war in Gaza, many pro-Palestinian activists have complained about instances of shadow banning of the content they post on social media. A similar situation occurred with content related to Black Lives Matter in 2020.[24]

71. This practice is hardly consistent with freedom of expression and information as it impacts severely on users' right to defend themselves and their possibility to modify the content in question so that it abides by the service's T&Cs. It also raises questions of fairness and data protection. The problem with shadow banning, however, is that it is difficult to prove that it actually happened.[25]

72. Instances of shadow banning have already been qualified as illegal by European courts. For example, on 3 June 2024, a Belgian court ruled against Meta for placing a shadow ban of a politician's Facebook page.[26] According to Meta, the politician had repeatedly posted content that violated Facebook's rules on hate speech, support for dangerous individuals and hate organisations, and bullying and harassment. Regarding the right to freedom of expression, the court ruled that Meta's actions should in principle not be tested directly against the principles of legality, legitimacy and proportionality set in Article 10 ECHR. Thus, the appellant could not claim a violation of the right to freedom of expression. However, Meta did not act in accordance with the principle of good faith execution of agreements when imposing the shadow ban and failed to provide sufficient procedural safeguards to users when applying content moderation measures. Meta also failed to provide sufficient justification for this decision.

73. One month later, on 5 July 2024, a Dutch court ruled against X (formerly known as Twitter) for shadow banning a user's account.[27] The appellant had made a post criticising the European Commission for spreading misleading information concerning its proposal for a regulation laying down rules to prevent and combat child sexual abuse. According to X, the restrictions may have been triggered by X's automated systems, and in January 2024, the applicant was informed that his post had been wrongfully associated with child sexual exploitation and that the restriction had been lifted. The interesting part of this case concerns X's arguing that its T&Cs reserved its right to limit access to various aspects and functionalities of its service, and that since the applicant had access to other key functionalities, X's obligations towards the user were fully met. The court found that the clause enabling Twitter to suspend or terminate access to its paid service at any time without any reason was contrary to the EU's Directive on unfair terms in consumer contracts. It further ruled that X had acted in breach of Article 12 and 17 DSA since its first two responses to the applicant's request for information were too vague and did not elucidate the exact reasons behind the restriction, and X's Help Centre did not enable effective communication between the company and its users.

7. Improving content moderation

74. To avoid content moderation practice which may violate users' rights, I would propose four directions on which public authorities and social media companies could work together: directly incorporating principles of fundamental rights law into T&Cs; enhancing T&Cs to ensure that they are clear; giving human moderators

---

[24] See These TikTok Creators Say They're Still Being Suppressed for Posting Black Lives Matter Content.

[25] See Social media users say their Palestine content is being shadow banned – here's how to know if it's happening to you, Meta's Broken Promises - Systemic Censorship of Palestine Content on Instagram and Facebook, and People are accusing Instagram of shadowbanning content about Palestine.

[26] Both anonymised judgments of the Court of Appeal in this case are available at: https://www.rechtbanken-tribunaux.be/sites/default/files/media/hbca/gent/files/uittreksel-inhoud-tussenarrest-24.10.2022-hof-van-beroep-gent-k-7.pdf and https://www.rechtbanken-tribunaux.be/sites/default/files/media/hbca/gent/files/uittreksel-inhoud-arrest-03.06.2024-hof-van-beroep-gent-k-7.pdf. See also Belgian far-right MEP wins case against Meta over shadowban.

[27] See both judgments of the District Court of Amsterdam at https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBAMS:2024:3980 and https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBAMS:2024:4019. See also [NL] District Court of Amsterdam rules that X has violated the DSA and the GDPR by "shadowbanning" its user.

appropriate training and good working conditions; and using AI in a more efficient way. This chapter concludes with an analysis of the role of recommender algorithms in promoting freedom of expression and fostering plurality and diversity of views.

7.1. Incorporating principles of fundamental rights law into T&Cs

75. According to the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression,[28] T&Cs should move away from "a discretionary approach rooted in generic and self-serving "community" needs". They should be rethought to allow users "to develop opinions, express themselves freely and access information of all kinds in a manner consistent with human rights law". Companies should incorporate into their T&Cs principles of human rights law that ensure that content moderation "will be guided by the same standards of legality, necessity and legitimacy that bind State regulation of expression".

76. As mentioned above, Article 14 DSA states that providers of intermediary services must act with due regard to the rights and legitimate interests of all parties involved, including users' fundamental rights as enshrined in the Charter of Fundamental Rights of the EU. A human-rights-based approach is also the one which the Council of Europe bodies foster.

77. Enshrining fundamental rights in T&Cs will strengthen their enforceability. However, the problem of balancing the freedom of social media and the freedom of expression of users remains.[29] In this respect, an effort should be made by social media companies, which should clarify how their content moderation policies are intended to reconcile these two freedoms when they may be in conflict.

7.2. Clarity of T&Cs

78. T&Cs must explain what content is allowed on the service and what are the consequences of posting content that is illegal or contrary to the social media company's rules. This explanation must be complete, accurate and given in a way that any user (and not only tech-savvy ones) may understand.

79. On 9 August 2023, the UK's communications regulator Ofcom published a report on the user policies of video-sharing platforms (VSPs), which shines a light on the approaches to designing and implementing T&Cs to protect users of six VSPs (BitChute, Brand New Tube, OnlyFans, Snap, TikTok and Twitch), and highlights examples of good practice. The key recommendations are the following:

- T&Cs should be clear and easy to understand:
  - T&Cs must be easily found and accessed – for services with large numbers of child users, this could mean having a separate section explaining how children are protected on the platform;
  - they are drafted in terms which can be understood by as many users as possible, including children and people who do not have advanced reading skills; and
  - techniques are studied to measure and improve user engagement with and understanding of T&Cs.

- T&Cs must protect all users from harmful material: Examples of good practice include:
  - covering the broad range of different types of restricted material that are likely to cause harm to children;
  - clarifying what content is and is not allowed in a way that children can understand; and
  - where maturity or sensitivity ratings are used, clearly explaining to users what sorts of content should be rated as inappropriate for children.

- T&Cs should be clear about what content is not allowed and what happens when rules are broken: This would include:
  - setting out what content is and is not allowed on the platform (unless exceptional reasons apply for not doing so); and
  - explaining all potential actions that could be taken by the service provider if a user breaks the VSP's rules.

---

[28] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 6 April 2018.
[29] See e.g. Using Terms and Conditions to apply Fundamental Rights to Content Moderation.

7.3. Appropriate training and good working conditions for human moderators

80. Every social media company uses (to a greater or lesser extent)[30] human moderators to take decisions on user-generated content.

81. Human moderators need appropriate training to acquire adequate knowledge of both the law and the company's guidelines they must apply. The Ofcom report mentioned above recommends setting out comprehensive definitions of key terminology, illustrated with case studies, as well as examples of context, exceptions, and considerations regarding harmful material. It further suggests providing audio and/or visual case studies of harmful content and providing detailed guidance on how moderators should respond to evolving harms and behaviours online that emerge in a crisis context. Moreover, T&Cs and guidance for moderators should be kept under review, and its effectiveness tested.

82. Human moderators also need good working conditions including (very importantly) mental healthcare. Problems related to working conditions include low wages, subcontracting, non-disclosure agreements (NDAs) that may prevent moderators from speaking up about the content they moderate and its effects on their mental health, and post-traumatic stress disorder (PTSD) disclaimers that must be signed by moderators acknowledging the mental health risks of the job.[31]

83. In Germany, a 2023 Manifesto by social media content moderators called for an end of exploitation in content moderation. The signatories demanded immediate, industry-wide change, notably through: better pay and benefit packages; a hazard bonus of at least 35% of moderators' annual salary; the provision of proper mental health care on a 24-hour basis; an end to culture of secrecy and bullying, including the dissolution of previously signed NDAs; the encouragement by social media companies for content moderators to collectively organise, bargain and join a union; putting an end to the outsourcing of critical safety work of content moderation; scrapping of all oppressive and unreasonable surveillance and algorithmic management; and equality of compensation irrespective of background or country of residence.

84. Regarding the mental health of human moderators, a 2023 study on the psychological impacts of content moderation on content moderators concluded that human moderators exposed to child sexual abuse material manifested with a range of symptoms that fit into a framework of post traumatic[32] and secondary traumatic stress[33] comparable to professionals working in the emergency services or caring professions, such as social workers. The study suggests that companies employing moderators should learn from these professions and provide psychoeducation and trauma-informed care to moderators.

85. In this regard, it is important that mental health problems caused by working as a content moderator are recognised as work-related illnesses and that employers and the judiciary are aware of this. In recent years, this issue has gained momentum, with human moderators willing to speak out in spite of NDAs and internal pressure. For example, in 2020 Facebook agreed to pay USD 52 million to settle a class-action lawsuit brought by moderators regarding mental health issues developed on the job. More recently, in January 2024 a Barcelona court ruled that the psychological problems of a content moderator working for a Meta subcontractor were a work-related illness. In this case, the court concluded that "the work stressor [was] the sole, exclusive and undoubted trigger" of the psychological damage caused to the employee, who was exposed to distressful content such as "self-mutilations, beheadings of civilians killed by terrorist groups, torture inflicted on people or suicides".[34]

7.4. Using AI in an efficient way

86. One of the main problems of moderating content on the Internet is its huge volume. According to World Economic Forum estimations, by 2025, about 463 exabytes will be created every day.[35] In this regard, AI can help substantially in automatising content moderation on online services. AI systems can quickly analyse and

---

[30] See Content moderation: Key facts to learn from Facebook, Instagram, X and TikTok transparency reports.
[31] See Facebook moderator: 'Every day was a nightmare' and Facebook and YouTube moderators sign PTSD disclosure.
[32] See Post-traumatic stress disorder (PTSD).
[33] See Secondary Trauma explained.
[34] See Primera sentencia contra la subcontrata de Meta en Barcelona por daños mentales a un empleado.
[35] See How much data is generated each day?

classify large amounts of content (including live content) in a way that humans cannot. They can also filter out the most disturbing content so that human moderators are not exposed to it. They can operate autonomously or be combined with human intervention for cases where the machine alone cannot do the job.

87. A 2019 Cambridge Consultants study explains the three different ways in which AI technologies may have a significant impact on content moderation workflows:

- AI can flag content for review by humans, increasing moderation accuracy;
- AI can create training data to improve pre-moderation performance;
- AI can assist human moderators by increasing their productivity and reducing the potentially harmful effects of content moderation on individual moderators.

88. However, AI content moderation is not a perfect solution. For example, a 2022 US Federal Trade Commission Report to Congress urged policymakers to be "very cautious" about relying on it as a policy solution, as AI tools can be inaccurate, biased, discriminatory by design, and incentivise reliance on increasingly invasive forms of commercial surveillance. Very importantly, they have difficulties in understanding context (humour, sarcasm, cultural references), which can lead to mistaken interpretations of the appropriateness of users' content, and they have to be trained continuously to adapt to changing forms of harmful content.[36]

89. Another important issue is the moderation of AI-generated content. AI technologies make it possible to create all kinds of media content (not only text, but also audio and video) that are virtually indistinguishable from reality. Therefore, users need to be informed when they are confronted with text, images or sounds that are AI-generated, as this type of content can be highly misleading and contain disinformation and hate speech, among other dangers.

90. Article 8 of the Council of Europe Framework Convention on artificial intelligence and human rights, democracy, and the rule of law requires that parties to the Convention "adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place in respect of activities within the lifecycle of artificial intelligence systems, including with regard to the identification of content generated by artificial intelligence systems".

91. Article 50(2) of the EU AI Act provides that synthetic audio, image, video or text content generated by AI systems is marked in a machine-readable format and detectable as artificially generated or manipulated. Providers must ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, considering the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art. This obligation does, however, not apply to AI systems that perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or were authorised by law to detect, prevent, investigate or prosecute criminal offences.

92. In September 2024, Meta's Oversight Board published the results of an investigation into the enforcement of content policies by AI algorithms and automation techniques. The investigation yielded a set of recommendations, which are outlined below:

- Social media should focus their policies on identifying lack of consent among those targeted by deepfakes. AI generation or manipulation should be considered as a signal that such images could be non-consensual.
- Automation should facilitate a more nuanced comprehension of policies and prevent the inadvertent removal of content created by users, including through the provision of informative notifications. Furthermore, users should be afforded the opportunity to provide context regarding their content that may not have been correctly interpreted by content moderators, whether human or AI. This could include instances where the content in question is intended to be satirical, raises awareness, or expresses condemnation.
- The benefits of new generative AI models should be shared equitably by social media companies' global user bases – beyond English-speaking countries or markets in the West.

[36] See The Impact of AI and Machine Learning on Content Moderation: Advancements and Challenges.

- Automated content moderation systems should undergo rigorous and continuous evaluation in order to assess their performance for users who are most vulnerable and at the greatest risk. New models should not intensify existing societal biases that may have a detrimental impact on marginalised groups and other individuals when they are introduced.
- Global human rights, freedom of expression and ethics experts should be consulted before deploying new AI content moderation tools. Their recommendations on risk mitigations and other safeguards should be incorporated into their design.
- Third-party researchers should be given access to data allowing them to assess the impact of algorithmic content moderation, feed curation and AI tools for user-generated content.
- Social media should implement labelling to alert users when content has been significantly altered and may potentially mislead. Furthermore, it is essential to allocate sufficient resources to human review that supports this process.

93. Regarding ways for informing users about content that is AI-generated, there are different options including content labelling,[37] using automated fact-checking tools,[38] forensic analysis,[39] and, especially, watermarking techniques.

94. Watermarking is a process of embedding into the output of an artificial intelligence model a recognisable and unique signal (i.e. the watermark) that serves to identify the content as AI-generated.[40] Watermarking has a number of benefits: it allows content authentication and data monitoring, indicating authorship and protecting copyright, and preventing the spread of AI-generated misinformation. So far, however, it is marred by a number of limitations and drawbacks, among them the following:

- Lack of interoperability of different watermarking systems;
- Technical difficulties in watermarking text-based content;
- Watermarks can be manipulated, removed, or altered.

7.5. The role of recommender systems

95. As recalled by Recital 70 DSA, online platforms use algorithms to suggest, rank and prioritise information, distinguishing through text or other visual representations, or otherwise curating information provided by users. Such recommender systems have a significant impact on users' retrieval and interaction with information online, and play an important role in the amplification of certain messages, the viral dissemination of information and the stimulation of online behaviour. As such, it is fundamental that users are appropriately informed about how recommender systems impact the way information is displayed, and can influence how information is presented to them. As explained above, Article 27 DSA introduced a number of rules regarding the transparency of recommender systems.

96. The Assembly has already dealt with the issue of algorithmic transparency. In particular, its report intitled "Social media: social threads or threats to human rights?" (Doc. 14844) reminds that algorithmic selection leads to a lack of exposure to diverse sources of information, a phenomenon known as "filter bubble" or "echo chamber", and contributes to radicalisation and growing partisanship in society. Algorithms can, however, be designed and implemented to encourage plurality and diversity of views, attitudes and opinions. Ideally, companies should call on some outside evaluation and auditing in order to determine that their algorithms are not biased and foster plurality or diversity of facts, points of views and opinions. Even though there are no mechanisms to make this recommendation mandatory, a "Seal of Good Practices" could be awarded to internet operators whose algorithms are designed to foster the selection of plural content, thus enabling ideologically cross-cutting exposure.

7.6. Summing up

97. Content moderation is a complex issue that involves making split-second decisions about literally millions of pieces of content. In addition, the strong pressure on social media companies to stop illegal and harmful content and to cooperate with public authorities, for example in the fight against war propaganda, disinformation and hate speech, may lead them to be overly cautious in moderating content and to remove

[37] See Misinformation warning labels are widely effective: A review of warning effects and their moderating features.
[38] See Emerging technologies and automated fact-checking: tools, techniques and algorithms.
[39] See Machine learning in digital forensics: a systematic literature review.
[40] See Generative AI and watermarking.

items that are legal. It is therefore crucial that any regulatory intervention in this domain does not result in unintended consequences for freedom of expression while duly considering the rights and interests of social media companies.

98. Building on these conclusions, I propose a set of concrete measures in the draft resolution.