ARTICLE 19

unesco
In partnership with UNESCO

Funded by the European Union

Content moderation and freedom of expression handbook

August 2023

SOCIAL MEDIA 4PEACE

ARTICLE 19

T: +44 20-7324 2500

F: +44 20-7490 0566

E: info@article19.org

W: www.article19.org

Tw: @article19org

Fb: facebook.com/article19org

This publication was produced with the financial support of the **European Union** and **UNESCO**.

Contents

Introduction

This publication has been produced as part of the United Nations Educational, Scientific and Cultural Organization's (UNESCO's) project **Social Media 4 Peace** funded by the European Union (EU).

Power of social media companies over freedom of expression

In their early days, social media platforms were widely seen as a powerful force for good, liberating free expression, enabling connections between people, and spearheading a democratic revolution across the world. This perception has now changed. Today, a very small number of large social media companies act as gatekeepers, controlling what a huge number of people get to see or say online. They have a direct impact on the dynamic of content distribution, as well as on online media diversity and freedom of expression.

This significant power and influence is coupled with the fact that the business models of the largest social media companies are often based on the collection of vast amounts of data about their users and their online habits (behavioural data) and the monetisation of this data through online (targeted) advertising. This significantly interferes with users' right to privacy and can have a negative effect on freedom of expression. Of particular concern is the spread of 'hate speech' and 'disinformation' on online platforms.[1] Social media companies have been accused of prioritising profit over user safety by using algorithms that promote the consumption of harmful content, including 'hate speech' and 'disinformation'. There are increased calls for social media companies to step up their content moderation efforts and tackle such problematic content.

Purpose and structure of this handbook

Content moderation includes the different sets of measures and tools that social media companies use to deal with content on their platforms that is either illegal or in violation of the companies' own community standards. It is influenced by a series of factors, ranging from the implementation of social media companies' business models, to pressure from

advertisers and regulators to avoid socially undesirable, harmful, or illegal speech, to the need to protect freedom of expression online.

Focusing on the largest social media platforms, this handbook provides a concise overview of the current state of content moderation and some of the main issues it raises from a freedom of expression perspective.

This handbook is structured as follows:

- First, it outlines the applicable standards for the protection of freedom of expression online that apply to content moderation, with a specific focus on 'hate speech' and 'disinformation'.
- Second, it addresses the contractual relationship between users and the largest social media companies through terms of service and community standards that govern online speech, and the issues raised by regulating speech by contract.
- Third, it explains the role of regulatory frameworks in content moderation, in particular the concept of intermediary liability and the recent trend towards greater regulation of social media companies.
- Finally, it discusses content moderation processes typically applied by the largest social media companies – including automated systems, human reviewers, and user and third-party reporting – with a special focus on the shortcomings of automated content moderation systems.

For an analysis of the specific challenges surrounding the disconnect between the largest social media companies' content moderation practices and the local communities where the moderated content is produced and distributed – based on a study of current practices in Bosnia and Herzegovina, Kenya, and Indonesia – see ARTICLE 19's report *Content Moderation and Freedom of Expression: Bridging the Gap between Social Media and Local Civil Society*.

About the project

This handbook is part of the **Social Media 4 Peace** project that UNESCO and ARTICLE 19 are implementing in Bosnia and Herzegovina, Kenya, Indonesia, and Colombia with the support of the EU. The overall objective of the project is to strengthen the resilience of societies to potentially harmful content spread online, in particular 'hate speech' and 'disinformation', while protecting freedom of expression and contributing to the promotion of peace narratives through digital technologies, notably social media. ARTICLE 19's contribution to the project focuses on concerns raised by current content moderation practices on the largest social media platforms in the four target countries.

Applicable international human rights standards

Overview

How do international human rights and freedom of expression standards apply to content moderation? Content moderation is influenced and shaped mainly by the actions of social media companies on the one hand and state actors, in particular through laws and regulations, on the other. Both have responsibilities under international human rights law, including when it comes to protecting freedom of expression, albeit to varying degrees.

This section will first outline the applicable standards for protecting freedom of expression online that should guide any measures adopted by states and social media companies in content moderation. Second, it will briefly provide an overview of how international freedom of expression standards apply to 'hate speech' and 'disinformation' – two categories of speech which capture a wide range of expression but lack a uniform definition under international human rights law. Third, it will explain the extent of human rights responsibilities for social media companies, how they differ from those of states, and what this means in practice for content moderation.

Guarantees to the right to freedom of expression

The right to freedom of expression is protected by Article 19 of the Universal Declaration of Human Rights (UDHR),[2] and given legal force through Article 19 of the International Covenant on Civil and Political Rights (ICCPR)[3] and in the regional treaties.[4]

The scope of the right to freedom of expression is broad. It requires states to guarantee to all people the freedom to seek, receive, or impart information or ideas of any kind, regardless of frontiers, through any media of a person's choice. States have an obligation not to interfere in the circulation of information and ideas or unduly restrict expression. States also have a positive obligation to promote conditions that are conducive to freedom of expression and protect individuals from disproportionate interference by private entities.[5] In this context, it has been argued that states are required to take positive measures to ensure that the right to freedom of expression can be effectively enjoyed

online, for example by introducing procedural safeguards in the legal framework regarding removal of online content.[6]

In 2011, the UN Human Rights Committee, the treaty body monitoring states' compliance with the ICCPR, clarified that the right to freedom of expression applies also to all forms of electronic and internet-based modes of expression.[7]

The UNESCO *Global Toolkit for Judicial Actors: International Legal Standards on Freedom of Expression, Access to Information and Safety of Journalists* offers a specific module concerning the new challenges of protecting freedom of expression on the internet.

Limitations on the right to freedom of expression

Under international human rights standards, states may, exceptionally, limit the right to freedom of expression, provided that such limitations conform to the strict requirements of the Three-Part Test under Article 19(3) of the ICCPR.[8] This requires that limitations must be:

- **Prescribed by law**
    - Any restriction must be formulated with sufficient precision to enable individuals to regulate their conduct accordingly. Overbroad restrictions are not allowed.
    - For example, a legal text criminalising the 'spreading of rumours in a way that is likely to affect the general wellbeing of the public' would be open to many different interpretations – including the meanings of 'rumour', 'spreading', or 'general wellbeing' – and would not meet the standard of quality required by the Three-Part Test.
- **In pursuit of a legitimate aim**
    - Restrictions are only permitted for (a) the respect of the rights or reputation of others and (b) the protection of national security or public order, or of public health or morals.

- For example, authorities are not allowed to place restrictions on the exercise of the right to freedom of expression for the purposes of ensuring respect for 'recognised religious text' or protecting religions from ridicule.

- **Necessary and proportionate in a democratic society**

  - Restrictions must demonstrate a direct and immediate connection between the expression and the protected interest. Also, if a less intrusive measure is capable of achieving the same purpose as a more restrictive one, the less restrictive measure must be applied.

  - For example, imposing a prison sentence for defamation would constitute a disproportionate restriction on freedom of expression. Reparation for defamation should instead be provided through civil law remedies or through alternative measures, including apologies, corrections, and the use of the right of reply. These can effectively address any harm to reputation without exerting a chilling effect on freedom of expression. The sanction of imprisonment is too severe for conduct that involves damage to a person's reputation. In general, the sanction of imprisonment should be reserved for the worst speech offences, such as incitement to genocide.

The Three-Part Test can be applied to all measures taken by the state, including legislative measures, policies, or judgments rendered against individuals. With regard to content moderation, the Three-Part Test would typically be applied to assess regulatory frameworks that govern social media companies or requests made by states to access user data or to restrict content.

As will be explained in more detail below, social media companies also have responsibilities to respect human rights, including freedom of expression, and should ensure that their products and services are in line with international human rights standards. Hence, the Three-Part Test might also be used to analyse whether companies' terms of service or individual content moderation decisions are in line with international freedom of expression standards.

Freedom of expression and 'hate speech'

Addressing in detail the different types of 'hate speech' and how they should be dealt with within a human rights framework goes beyond the scope of this handbook.[9] This section therefore merely outlines a number of basic distinctions and principles when it comes to permissible restrictions of 'hate speech' under international freedom of expression standards.

There is no agreed definition of 'hate speech' in international human rights law. Put simply, 'hate speech' is any expression of discriminatory hate towards people. And it is the protection of equality and the principle of non-discrimination, as also protected in the ICCPR, which motivates most responses against 'hate speech'. The principle of non-discrimination protects individuals from any distinction, exclusion, or restriction based on a protected characteristic. Some of these are listed in Article 26 of the ICCPR; they may include race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth, or other status.

However, 'hate speech' defined broadly as any expression of discriminatory hate towards people does not necessarily entail a particular consequence. This lowest-common-denominator definition captures a very broad range of expression, including lawful expression. This definition, therefore, is too vague for use in identifying expression that may legitimately be restricted under international human rights law.

ARTICLE 19 therefore proposes to divide 'hate speech' into three categories, as follows.

'Hate speech' that must be prohibited

International criminal law and Article 20(2) of the ICCPR require states to prohibit certain severe forms of 'hate speech', including through criminal, civil, and administrative measures.

Article 20(2) of the ICCPR requires states to prohibit by law 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence' (incitement).

The Rabat Plan of Action, which provides authoritative guidance to states on implementing their obligations under Article 20(2) of the ICCPR,[10] outlines a six-part threshold test for expressions considered criminal offences under Article 20 of the ICCPR. It requires taking into account (1) the social and political context, (2) the status of the speaker, (3) intent to incite the audience against a target group, (4) the content and form of the speech, (5) the extent of its dissemination, and (6) the likelihood of harm, including imminence.

Hypothetical case

In the run-up to a heavily contested presidential election, the incumbent President makes a series of speeches to large rallies. During these rallies, he promulgates a rumour that supporters of the opposition, mostly belonging to another ethnic group, are arming themselves and are an existential threat to his supporters. As tensions increase, he uses racialised language, evoking instructions used in mass killings in the country a few decades earlier, calling on his supporters to take urgent action to secure an election victory.

Here, the President has engaged in 'hate speech' which would arguably reach the threshold of advocacy of hatred that constitutes incitement to violence. He understands and is exploiting ethnic tensions in society, and he knows as an influential politician that his use of a particular term would be understood and likely acted on violently by individuals in the crowd against members of the ethnic group associated with the opposition.

'Hate speech' that may be prohibited

States may prohibit other forms of 'hate speech', provided they comply with the requirements of Article 19(3) of the ICCPR. These would include individually targeted forms of bias-motivated threats, assault, or harassment.

Hypothetical case

A same-sex couple, both women, are confronted on a train by another passenger who starts shouting sexist and homophobic abuse at them, causing the pair to reasonably fear immediate physical violence.

In many jurisdictions, this incident would, appropriately, be prosecuted as a bias-motivated crime. The abusive passenger's expressive act falls within our broad typology of 'hate speech' and also amounts to the crime of assault. The credible threat of violence in the expression makes it criminal conduct, and since it is characterised by bias, the content of the expression is also evidence of bias motivation.

Lawful 'hate speech'

This should be protected from restriction under Article 19(2) of the ICCPR, but nevertheless raises concerns in terms of intolerance and discrimination and merits a critical response by the state.

Hypothetical case

A teenage boy, with a small number of followers on Twitter, tweets an offensive and sexist joke that trivialises the disappearance and likely murder of a local schoolgirl. It provokes a strong critical response against the boy online, and he eventually deletes the tweet. Though the communication is offensive and reflects a broader problem of misogyny in society, he did not intend to incite any harmful conduct against a particular group, and in any case, he does not have this kind of influence over his followers. This kind of 'hate speech' may justify soft intervention from local actors in positions of authority, such as teachers in the boy's school or other community leaders, but it does not justify the state imposing sanctions or other restrictions.

These distinct categories of 'hate speech' are important to keep in mind when assessing restrictions on freedom of expression, in particular those imposed by states, including when they relate to content moderation issues.

As explained in more detail below, social media companies often remove some content that is protected under international freedom of expression standards. Removal often occurs through enforcement of platforms' 'hate speech' and related policies. TikTok, for example, does not allow 'any hateful ideologies', in which it includes 'misogyny'. A blank prohibition of misogyny would not meet the requirements under Articles 19 and 20 of the ICCPR. TikTok also does not allow 'denying well-documented historical events that harmed groups based on a protected attribute' and states that it provides 'some protections related to age'.[11]

Policies can vary quite significantly across platforms. For example, while TikTok includes 'gender' as well as 'gender identity' in its list of protected characteristics, Meta lists only 'gender identity'.

While social media companies, as private entities, are entitled to adopt community standards which are stricter than the requirements set out earlier, their terms of service and content moderation decisions should at least be in line with international human rights norms and principles.

Example

In the 'Russian poem' case, the Meta Oversight Board – a mechanism established by Meta to review content moderation decisions in select cases and provide guidance on Meta's wider content moderation policies through policy advisory opinions – assessed Meta's decision to remove an April 2022 Facebook post. The post was published following Russia's unlawful invasion of Ukraine and compared the Russian army in Ukraine to Nazis, quoting a poem that calls for the killing of fascists. The Board stated that '[i]n order to assess the risks posed by violent or hateful content, the [Oversight] Board is typically guided by the six-factor test described in the Rabat Plan of Action, which addresses advocacy of national, racial or religious hatred that constitutes incitement to hostility, discrimination or violence'.

In this case, the Oversight Board found that despite the context of ongoing armed conflict between Russia and Ukraine and the charged cultural references employed by the user, it was unlikely that the post – a warning against a cycle of violence – would lead to harm. The Oversight Board concluded that the initial content removal was not necessary.

Freedom of expression and 'disinformation'

Like 'hate speech', the concepts of 'disinformation', 'misinformation', propaganda, and 'false information' do not have an agreed definition in international or regional human rights law.[12] In general, attempts to define these concepts in national laws and regional standards focus on prohibitions of 'false' or 'misleading' information that may cause certain 'harm' or detriment.[13]

As restrictions on 'disinformation' engage limitations on freedom of expression, any legal and policy response to these problems must be based on international freedom of expression standards. In particular, it must meet the Three-Part Test under Article 19(3) of the ICCPR. Restrictions on 'disinformation' are allowed only if they are explicitly linked to protecting the legitimate aims specified under Article 19(3) – such as the respect of the rights or reputation of others and the protection of national security, public order, or public health or morals – or Article 20 of the ICCPR – prohibiting propaganda for war and any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence. The mere falsity or misleading nature of certain information is not sufficient to restrict its dissemination. In addition, for restrictions on 'disinformation' to be compatible with international standards on freedom of expression, they need to be applied in a proportionate manner.

In general, the harmful consequences of 'disinformation' should be dealt with through enabling measures, such as ensuring a free, independent, and diverse media environment and promoting individuals' exposure to the broadest possible diversity of information. Bans and other legal restrictions on the sharing of false information, on the other hand, are open to abuse and can have a devastating impact on political discourse. Enacting a legal duty of 'truth' creates a powerful instrument to control the flows of information and ideas, which can be a dangerous tool in the hands of public authorities. And indeed, the concepts of 'disinformation', 'misinformation', propaganda, and 'false information' have been used and abused by power-holders as a means of cracking down on dissent and discredit information that they do not like.

Although legislation criminalising the dissemination of 'fake news' in some form is nothing new, there has been a lot of regulatory activity in recent years. Governments around the world have introduced or updated 'disinformation' offences, including through legislation purportedly combating cybercrimes, adopted in the context of the Covid-19 pandemic, or aimed at suppressing reporting on armed conflicts.[14] This, and similar legislation, has been used to arrest and prosecute bloggers, journalists, and critics of governments.

Example

In November 2022, Senegalese journalist Pape Alé Niang was arrested and detained by Senegalese authorities. He was charged with, among other things, 'dissemination of false news likely to discredit public institutions'. The charges stemmed from his coverage of rape allegations against Senegal's main opposition leader, Ousmane Sonko, which had created political tensions in the country. Pape Alé Niang was eventually released on 10 January 2023 but was placed under judicial supervision. Pape Alé Niang is one of several journalists reporting on matters of public interest who have recently been prosecuted under laws criminalising the dissemination of 'false information'.

Example

In December 2022, one of Russia's most prominent opposition figures, Ilya Yashin, was jailed for eight-and-a-half years for condemning the killing of hundreds of Ukrainian civilians by Russia's occupying forces in Bucha on his YouTube channel. In this video, he shared images and stories from the scene by the BBC and others. The court found that he had knowingly disseminated false information about the Russian armed forces.

The conviction was based on a Russian law introduced following its illegal invasion of Ukraine providing for prison terms of up to 15 years for those convicted of disseminating 'fake news' or any information that Russian authorities deemed to be false in war-related coverage.

It is problematic not only if public authorities become arbiters of truth, but also if private entities that control quasi-public spaces of speech get to decide what information is 'correct' and what is 'false' and to suppress and restrict information they deem to be incorrect. State regulation requiring platforms to take measures to suppress false information or terms of service containing blanket bans of 'disinformation' would therefore constitute undue interferences with the freedom of expression rights of platform users.[15]

Human rights responsibilities of social media companies

States are the primary duty bearers under international human rights law. They have a duty to respect, protect, and fulfil human rights. This includes a duty to protect individuals against human rights abuses by all actors in society, including businesses. Indeed, states must prevent, investigate, punish, and redress human rights abuses by private actors. Although some of the largest social media companies – just like other large enterprises – arguably have more power than certain states and can profoundly impact the human rights of individuals and communities wherever they operate, as businesses they do not have the same level of human rights obligations as states.

However, the *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework* (the Guiding Principles) recognise that companies have a responsibility to respect human rights, independent of state obligations or the implementation of those obligations.[16]

In particular, the Guiding Principles recommend that companies should:[17]

- make a public statement of their commitment to respect human rights, endorsed by senior or executive-level management;
- conduct due diligence and human rights impact assessments to identify, prevent, and mitigate any potential negative human rights impacts of their operations;
- incorporate human rights safeguards by design to mitigate adverse impacts, and act collectively to strengthen their leverage vis- à-vis government authorities;
- track and communicate performance, risks, and government demands; and
- make remedies available where adverse human rights impacts are created.

Social media companies more specifically may have to assess and mitigate potentially negative human rights impacts of, among other things, their response to government takedown or data access requests or their own data collection practices, content curation,

and advertisement targeting systems. Where a social media company contributes to human rights violations, they should provide an effective remedy to affected communities.

When it comes to content moderation practices, social media companies should, among other things, establish clear and unambiguous terms of service in line with international human rights norms and principles,[18] produce transparency reports about government demands,[19] ensure that sanctions for non-compliance with their terms of service are proportionate, and provide effective remedies for affected users in case of violations.[20]

Civil society has made specific recommendations on social media companies' responsibility to respect international human rights standards. For instance, the Manila Principles on Intermediary Liability state that companies' content restriction practices should comply with the tests of necessity and proportionality under human rights law (Principle 4) and should provide users with complaint mechanisms to challenge companies' decisions (Principle 5(c)).

Some social media companies have also publicly committed to certain human rights principles, either through adopting a human rights policy, like Meta, or through membership in multi-stakeholder initiatives like the Global Network Initiative, whose members commit to collaborating in advancing user rights to freedom of expression and privacy.

To understand the extent to which the responsibility to respect international human rights law may guide individual content moderation decisions, it can be instructive to read decisions by Meta's Oversight Board, as its sources of authority are both Meta's content policy and Meta's human rights responsibilities in accordance with the UN Guiding Principles.

Example

The Oversight Board assessed a case concerning the removal of content posted by an Instagram user in January 2021 featuring an image of Abdullah Öcalan with a comment that it was time to talk about ending Öcalan's isolation in prison. Öcalan is one of the founder members of the Kurdistan Workers' Party (PKK), which is designated as a terrorist organisation in Turkey, and has been in prison since 1999. Both the PKK and Öcalan are designated as dangerous entities under Facebook's Dangerous Individuals and Organisations policy.

ARTICLE 19 provided public comments submitting that any restriction on freedom of expression on Facebook should be guided by human rights standards, that policies providing for restrictions of content on the basis that it might incite terrorist activity need to operate with narrowly defined terminology, and that there was a significant discrepancy between the human rights approach to 'terrorist content' and the Facebook community standards, which focused overly on speakers or organisations.

The Oversight Board found that removing the post was inconsistent with Meta's commitment to respect human rights, because discussing the conditions of any individual's detention constituted protected speech; the community standards were not sufficiently clear to meet the legality test; and the removal was unnecessary and disproportionate given that the content in this case did not indicate any support for violent acts committed by Öcalan or by the PKK.

There appears to be limited case law by judicial bodies addressing the role played by human rights when it comes to content moderation decisions. One notable exception is the judgment issued by the Court of Rome in *Facebook v. CasaPound*.

Example

The Court of Rome dismissed Meta's (then Facebook's) appeal against a preliminary injunction ordering it to reactivate the account of the Italian far-right party CasaPound. Meta had deactivated the account without notice or explanation and argued before the court that its actions were legitimate on the grounds that the account included content which constituted 'hate speech' and incitement to violence, in violation of Meta's Terms of Use.

The Court sided with CasaPound and held that the contract concluded between Meta, although an ordinary civil law contract, should be interpreted in accordance with the Italian Constitution, including the right to free expression, as Meta held a de facto systemic role for the purposes of political participation.

The Court rejected Meta's argument that the ban 'sanctioned' the fact that CasaPound was a political organisation intrinsically against the Constitution and human rights law. It held that it was not up to Meta to determine whether CasaPound was a legitimate actor, also taking into consideration that CasaPound had not been outlawed by the competent Italian authorities. The Court found further that the contractual relationship in this case was unduly terminated and that the contents shared by CasaPound did not reach a degree of gravity such as to justify an outright ban.

In a similar vein, the Federal Supreme Court of Germany found in 2021 that Meta (then Facebook) was bound by German fundamental rights (including the right to freedom of expression).

Example

Ruling on a case involving posts deleted and accounts suspended by Facebook due to allegations of 'hate speech', Germany's Federal Supreme Court held that the company had to strike a balance in its terms of service between conflicting fundamental rights, namely users' right to freedom of expression and Meta's right to exercise a profession guaranteed by German Basic Law. This required Meta to inform users at least retrospectively about post removals and in advance about account blockings, provide reasons for these actions, and allow users to respond, followed by a new decision.

Meta's terms and conditions for post removal and blocking did not meet these requirements. Consequently, the company was not entitled to delete the plaintiffs' posts or block their user accounts.

Social media companies often find themselves faced with legal demands from governments that may not be in line with international human rights standards – for example, to block certain accounts, take down content, or provide access to user data. Such demands may be based on cybercrime, 'disinformation', or other laws that do not meet international freedom of expression standards.

As explained previously, companies' responsibility to respect human rights is independent of a government's willingness to fulfil its human rights obligations. The former UN Special Rapporteur on Freedom of Expression has specified that companies' responsibility to respect freedom of expression should, as a minimum, include a duty to 'engage in prevention and mitigation strategies that respect principles of internationally recognised human rights to the greatest extent possible when faced with conflicting local law requirements'.[21] Legal demands should be interpreted and implemented as narrowly as

possible to ensure the least possible restriction on the right to freedom of expression.[22] The Special Rapporteur further held that when companies receive such requests, they should 'seek clarification or modification; solicit the assistance of civil society, peer companies, relevant government authorities, international and regional bodies and other stakeholders; and explore all legal options for challenge'.[23] Finally, companies must be transparent about government requests and provide details on the type of content subject to the requests (e.g. defamation, 'hate speech', terrorism-related content) and the actions taken by the companies (e.g. partial or full removal, country-specific or global removal, account suspension, removal granted under terms of service).[24]

Content moderation in practice

Key terminology

It is useful to briefly address how content moderation differs from content curation. Content moderation and content curation systems and processes can be closely connected, but they raise different issues and are often treated separately by regulators.

For the purposes of this handbook, we rely on the following definitions:

- **Content moderation** includes the different sets of measures and tools that social media platforms use to deal with illegal content and enforce their community standards over user-generated content on their service. This generally involves flagging by users, 'trusted flaggers', or 'filters'; removal, labelling, down-ranking, or demonetisation of content; or disabling certain features.

- **Content curation** is how social media platforms use automated systems – often referred to as recommendation systems – to rank, promote, or demote content in news feeds, usually based on their users' profiles. Content can also be promoted on platforms in exchange for payment. Platforms can also curate content by using interstitials – warning messages displayed before the content is shown – to caution users about sensitive content or by applying certain labels to highlight, for instance, whether the content comes from a trusted source.

Put simply, content moderation is about ensuring that the content published does not violate any rules. Content curation is concerned with how the content is prioritised and presented to a user, for example what appears at the top of a users' news feed. There may be overlap between these processes. For example, down-ranking a piece of content can be a content moderation measure but is also an inevitable part of the content curation process.

Terms of service and community standards

Sharing information or opinions on social media platforms is not control-free. When users join Facebook, TikTok, Twitter, or YouTube, they accept that they must abide by those companies' terms of service, which govern the contractual relationship between a social media company and the user. This contractual relationship sets the parameters of, for example, access to and use of the different products, apps, and services offered.

These terms of service typically include the platform's community standards, sometimes referred to as community guidelines or policies (see, for example, Facebook's Community Standards, the Twitter Rules and Policies, or YouTube's or TikTok's Community Guidelines). These community standards typically lay down the types of content that the company allows or prohibits on its platform. In addition, they may bar certain groups or individuals considered dangerous or prohibit certain online behaviours (such as impersonation or spam).

Social media users who fall foul of these standards may see their content removed or down-ranked, or their account may be disabled altogether.

In terms of the focus of this handbook on 'hate speech' and 'disinformation', most of the different companies' standards deal with this type of content in one way or another, albeit with varying degrees of precision and labelling the content in different ways.

For example, when it comes to community standards dealing with 'false' information in the broadest sense, the approach can vary quite significantly between platforms.

- TikTok's rules on 'integrity and authenticity' state that they remove content or accounts that involve misleading information that causes significant harm. TikTok further defines 'harmful misinformation' to be removed as content that is inaccurate or false and that causes significant harm to individuals, their community, or the larger public, regardless of intent. Significant harm may include the 'undermining of public trust in civic institutions and processes such as governments, elections, and scientific bodies',

but it does not include 'simply inaccurate information, myths, or commercial or reputational harm'.

- Twitter focuses its actions – which range from limiting amplification to removing or contextualising – on 'false' or 'misleading' information that could bring harm to populations affected by crises (such as situations of armed conflict, public health emergencies, and large-scale natural disasters); misleading media (defined as synthetic, manipulated, or out-of-context media that may deceive or confuse people); or content that is intended to manipulate or interfere in elections or other civic processes.

- Meta states that it removes 'misinformation where it is likely to directly contribute to the risk of imminent physical harm'. It further removes 'content that is likely to directly contribute to interference with the functioning of political processes and certain highly deceptive manipulated media'.

All of these platforms also have policies on behaviour in other areas that often overlap with the spread of 'misinformation', for instance on fake accounts, fraud, or coordinated inauthentic behaviour.

Each of these community standards could undergo a detailed analysis and raises separate issues and concerns from a freedom of expression perspective (see, for reference, ARTICLE 19's 2018 analysis of Facebook, Twitter, and YouTube's terms of service and community standards). While such analysis is beyond the scope of this handbook, we set out a number of general concerns with the current system in the next section.

Concerns raised by the regulation of speech by contract

The privatisation of speech regulation (i.e. the regulation of speech by contract) raises serious concerns for the protection of freedom of expression. These concerns are exacerbated by the fact that a small number of social media companies hold immense power over what people see and share online, with little public accountability.

Lower free speech standards

Community standards in the terms of service typically have lower standards for restrictions on freedom of expression than those permitted under international human rights law. For example, as mentioned in the context of companies' 'hate speech' policies, TikTok's policies do not allow misogyny. However, a blank prohibition of misogyny would not meet the requirements under Articles 19 and 20 ICCPR.

In addition, some social media companies reserve a right in their terms of service to remove any content at their sole discretion and without any reason. For example:

- Snapchat's Terms of Service provide: 'We may terminate or temporarily suspend your access to the Services if you fail to comply with these Terms, our Community Guidelines or the law, for any reason outside of our control, or for any reason, and without advanced notice'.

- Tiktok's Terms of Service state: 'We reserve the right, at any time and without prior notice, to remove or disable access to content at our discretion for any reason or no reason'.

ARTICLE 19 acknowledges that social media companies are in principle free to restrict content on the basis of freedom of contract, but they should still respect human rights, including the rights to freedom of expression, privacy, and due process in line with the Guiding Principles. Reserving the right to remove content 'for any reason or no reason' clearly falls short of the responsibility to respect human rights.

While social media companies legally have more leeway to restrict speech on their platforms compared with what states are allowed to restrict under international human rights law, it is also problematic that the content moderation rules in these quasi-public spaces are not guided by the principles of necessity and proportionality. Indeed, they often appear dictated by companies' desire to increase profit and meet the demands of the advertising industry, which does not want to be associated with offensive, shocking, or disturbing content, although such content enjoys protection under freedom of expression

standards. In practice, low free speech standards are also often the result of companies adapting their community standards to domestic legal requirements that fall below international standards on freedom of expression.[25]

Lack of transparency and accountability

There is a considerable lack of transparency around the implementation of community standards. This in turn negatively impacts the ability to hold companies accountable for wrongful, arbitrary, or discriminatory content takedowns.

Sometimes the lack of transparency can concern the content of a policy. Meta's Dangerous Organisation and Individuals policy has been a prominent example of this issue. Under this policy, Meta designates individuals, organisations, and networks of people that the company deems to 'proclaim a violent mission' or to be 'engaged in violence' and removes 'praise' or 'support' for such entities. However, while this policy accounts for a large amount of content takedown on Meta's platforms, the actual list of designated entities is not publicly available, making it impossible to scrutinise its content.

It is further problematic that community standards are not fully accessible in many languages, making it impossible for many users to understand the rules that govern their online speech.[26]

When it comes to transparency reporting, some social media companies have made efforts to improve their practices in recent years. However, their current reporting on how they enforce their community standards still lacks the level of detail and quality required to derive meaningful insights.

To provide just one illustrative example, Meta publishes information about content removals on the basis of its terms of service in its transparency reports, broken down by policy area. However, sharing aggregate data on how much content has been removed on the basis of such a complex and broad category as 'hate speech' is not particularly informative. It lacks information on, among other things, how Meta operationalises its definition of 'hate speech' (and what data it feeds its automated content moderation

tools); the number of removals broken down by different categories (e.g. by protected characteristics) and by country (to understand whether the community standards are applied differently from country to country); and what percentage of takedowns based on the 'hate speech' policy were made following notification by government agencies in the different countries. Most transparency reporting also focuses on removal of content only, while other measures like down-rating are not addressed.[27]

Such lack of transparency makes it generally difficult to know whether community standards are applied reasonably and consistently or in an arbitrary and discriminatory manner, unless there is press coverage of specific cases or public campaigns conducted by affected individuals or groups.

UNESCO's 2021 brief of accountability and transparency in the digital age, *Letting the Sun Shine In*, sets out a list of illustrative high-level transparency principles that could enhance the transparency of online platforms, focusing, among other things, on transparency of content and process, personal data gathering and use, and due diligence and redress.

Lack of procedural safeguards and remedy

There are insufficient procedural safeguards that apply to the removal of content on social media. Here again, transparency is an issue. It is not always clear whether companies notify users that their content has been removed or flagged, or whether their account has been penalised in any way, and the reasons for such actions. Even where notification is provided to the user, it often contains a simple reference to a policy allegedly violated, without sufficient reasoning for the user to understand why restrictive action has been taken. As explained in more detail below, this problem is exacerbated where automated content moderation tools are involved.

While most of the largest social media companies give users the ability to appeal against content takedowns or account suspensions, this is only meaningful if users are properly notified and understand the reasoning behind the sanctions that led to a restriction of their content. In this context, it is particularly problematic that social media companies have increasingly adopted a practice colloquially known as 'shadowbanning', which describes

instances where users have their content hidden or reduced in visibility without being informed by the platform.

Another critical shortfall is that individuals whose content is removed generally lack adequate legal remedies. Companies' terms of service will often not grant them any basis for claims relating to content restrictions, and users in most jurisdictions will also not be able to resort to non-contractual claims. Problematic dispute resolution and choice of law clauses in social media companies' terms of service can create additional barriers to access to justice for users. In some cases, they may bar users from bringing claims in the local courts of their countries of residence or from applying their local laws to the terms of service. This will deter most users from bringing litigation, as they will lack the resources to do so.

For example, for users based in the UK – and similar provisions will likely apply in other jurisdictions – Twitter provides that all disputes related to the terms of service have to be brought before the courts of San Francisco and that the laws of the State of California apply. Snapchat's terms of service contain an arbitration agreement that provides for arbitration in the US and contains a class action waiver (except for the use of applicable small claims court procedures and with a 30-day opt-out option that most users will never notice). Meta's terms are more reasonable, as they provide that for consumer disputes, the court in the state of the users' main residence will have jurisdiction and the laws of said state will apply. TikTok's Terms of Service also stipulate that disputes related to the Terms of Service are subject to the jurisdiction of the users' local courts, as well as the courts of the Republic of Ireland and the courts of England & Wales.

Circumventing the rule of law

Finally, public authorities, and law enforcement agencies in particular, regularly seek the cooperation of social media platforms with a view to combating criminal activity (e.g. dissemination of child sexual abuse material) or other social harms (e.g. 'online extremism') in a way which circumvents the rule of law. In particular, because these authorities do not always have the power to order the removal of the content at issue, they

sometimes contact social media companies informally and request the removal of content on the basis of the companies' terms of service. While companies will often not be legally required to comply with such requests, they are put in a difficult position, particularly in circumstances where the content may be at the fringes of illegality. The net result is that social media companies often become the long arm of the law without users being afforded the opportunity to challenge the legality of the restriction at issue before the courts.[28]

Example

In the 'UK drill music' case, the Oversight Board overturned Meta's decision to remove a UK drill music video clip – 'Secrets Not Safe' by Chinx (OS) – from Instagram. Meta originally removed the content following a request from the UK Metropolitan Police. The Metropolitan Police had emailed Meta requesting that the company review all content containing 'Secrets Not Safe' and had provided additional context to Meta, covering information on gang violence, including murders, in London and the police's concern that the track could lead to further retaliatory violence. Meta removed the content from the account under review for violating its violence and incitement policy.

The Oversight Board found, among other things, that '[t]he channels through which law enforcement makes requests to Meta are haphazard and opaque. Law enforcement agencies are not asked to meet minimum criteria to justify their requests, and interactions therefore lack consistency. The data Meta publishes on government requests is also incomplete.'

The role of regulatory frameworks

Content moderation processes deal not only with the enforcement of social media companies' terms of service and community standards but also with regulatory requirements to remove illegal or objectionable content. Companies have been subject to increased pressure from governments over the last few years to remove more content from their platforms – from 'hate speech' and 'extremist' content to 'disinformation'.

Traditional focus on governing intermediary liability

Traditionally, regulatory frameworks have focused on governing the liability of so-called internet intermediaries – a broad term which includes web hosting companies, internet service providers, search engines, and social media platforms.[29] Laws dealing with intermediary liability regulate the extent to which internet intermediaries can be held legally responsible – and whether they are required to, for example, pay monetary damages to an aggrieved party – for content disseminated or created by their users (third-party content).

Generally speaking, liability regimes range from strict liability at one end of the spectrum to immunity at the other. Under strict liability regimes, internet intermediaries can be sued in court for user misconduct, without the need for any fault or knowledge on the part of the intermediary. Intermediaries are effectively required to monitor content and take action where relevant in order to comply with the law. This model has been applied in Thailand, for example.

Example

Chiranuch Premchaiporn, the editor of Prachatai, an online news site in Thailand, was tried and convicted under the provisions of Thailand's Computer Crimes Act 2007 for failing to expeditiously remove an anonymous comment that was deemed insulting to the King.[30] The Computer Crimes Act punishes 'false data' that damages a third party, causes public panic, or undermines the country's security, and 'any service provider intentionally supporting' the false data. Thailand's criminal code (*lèse-majesté*) states that anyone who 'defames, insults or threatens the king, the queen, the heir-apparent or the regent' will be sentenced to prison.

Chiranuch Premchaiporn was sentenced to one year in prison with a fine of 30,000 baht, which was reduced to eight months with suspension and 20,000 baht fine after appeal.[31]

Blanket immunity from liability for user-generated content – which means any claims against intermediaries on the basis of user-generated content would be barred – is uncommon. The most prominent example that does endorse such an approach to a large extent is Section 230 of the 1996 Communications Decency Act, which applies in the United States. Section 230 grants legal immunity to online platforms for content posted by third parties (although it does not extend to immunity for violations of federal criminal law, intellectual property law, or electronic communications privacy law).

Many legal systems fall somewhere in between strict liability and blanket immunity through a system of conditional immunity. Often, internet intermediaries are immune from liability as long as they remove content once they obtain *actual knowledge* of illegality. Such knowledge-based liability systems usually operate via so-called 'notice and takedown' procedures.

Exactly how such 'notice and takedown' procedures work varies between jurisdictions. Typically, it is considered that an internet intermediary acquires knowledge of the illegal nature of a content once it is notified of it by a third party. If the internet intermediary does not remove the illegal content despite such notification, it may be held legally responsible for any damage caused. For example, the recently adopted EU Digital Services Act imposes liability for content that has been the subject of properly substantiated notices by users.

From a freedom of expression perspective, it is widely recognised – including by the special mandates on freedom of expression – that broad immunity from liability for internet intermediaries is one of the most effective ways of protecting free speech online. If companies can be held liable for the content published by its users, it effectively requires them to monitor all user-generated content – a massive invasion of users' privacy rights. Such a regime also provides a strong incentive for companies to over-censor their users and take down material that may be perfectly lawful to avoid any risk of breaching the law.

Experience shows that even conditional immunity regimes operating via 'notice and takedown' procedures provide an incentive to remove content promptly on the basis of

allegations made by a private party or public body, without a judicial determination of whether the content at issue is unlawful. Moreover, the person who published the content at issue is usually not given an opportunity to consider the complaint.

In 2011, Frank La Rue, the former UN Special Rapporteur on Freedom of Expression, stated that censorship should never be delegated to a private entity and that states should not use or force intermediaries to undertake censorship on their behalf.[32] He also noted that 'notice and takedown' regimes were subject to abuse by both states and private actors, and that the lack of transparency in relation to decision-making by intermediaries often obscured discriminatory practices or political pressure affecting companies' decisions.[33]

Trend towards greater regulation of online platforms

In recent years, social media companies have been increasingly criticised for growing profits on the back of algorithms that promote addictive engagement with 'extremist' and other 'harmful' content. This has raised the question of whether greater regulation is needed to tame the power of the largest social media companies, tackle illegal and other harmful content, and provide greater democratic accountability to the wider public for their decisions. Governments have responded with proposals under which platforms have a 'duty of care' to their users to prevent 'harm' caused by the speech of other users of the platform.[34]

Many of the current proposals for regulatory frameworks are deeply problematic from a freedom of expression perspective. While they ostensibly aim at increasing the accountability of social media companies, they often actually focus on online 'content' regulation. This means that they often regulate users' speech rather than the products, systems, and processes applied by the social media companies. States are effectively demanding that companies police human communications and decide what speech is 'illegal' or 'harmful',[35] when it should instead be the responsibility of independent judicial authorities to make such a determination.

In addition to the legitimacy concerns about outsourcing decisions on the legality of users' speech to private actors, in most cases these assessments are extremely complex and

context-dependent and should therefore be made by trained individuals. The reality is, however, that social media companies deploy algorithmic moderation systems, such as automated hash-matching and predictive machine-learning tools, to conduct content moderation. As these technologies are currently not advanced enough (and may never be) to distinguish legal from illegal content in a reliable manner, they routinely identify legal content as illegal and remove vast amounts of legitimate content.

Regulation should adopt a human rights-based approach and include digital markets

Instead of mandating platforms to restrict undesirable types of users' speech, proposals should ensure that human rights lie at the heart of platform regulation. The principles of legality, legitimacy, necessity, and proportionality set out in Article 19(3) of the ICCPR must be applied throughout. Like any framework that imposes limitations on free expression, regulation governing social media companies should be grounded in robust evidence and prioritise the least censorial and restrictive measures to address online harms.

What does this mean in practice? Instead of asking platforms to exercise even more powers over users' speech by screening and assessing all content they generate, regulators should focus on less intrusive methods that are specifically tailored to tackling some of the negative effects of social media companies' business models, including their recommendation systems. For example, regulatory solutions should require companies to be more transparent towards regulators, researchers, and users about how their recommendation systems work; set clear limits on the amount of user data that companies are allowed to collect; and mandate the performance of human rights due diligence. They should also focus on transparency about content moderation decisions and on improving systems to resolve any disputes these decisions cause.

ARTICLE 19 has also advocated for years that regulatory solutions should further address the dominant position of the biggest online platforms through regulatory tools that would increase competition in the market and enhance users' choice about what content they get to see online.

Some of these regulatory solutions were adopted in the EU in 2022 with the EU Digital Services Act and the Digital Markets Act. While these regulatory frameworks could have been more ambitious about protecting human rights online (e.g. by establishing an explicit right for users to encryption and anonymity), they rightly focus on rebalancing digital markets and regulating the content moderation and curation systems used by social media companies.

In addition, a number of the EU Digital Services Act's rules on content moderation tackle some of the concerns raised by the 'regulation of speech by contract'. They require, among other things:

- that users be notified and provided a statement of reasons in case of restrictions of the visibility of their content (including removal or demotion);
- the establishment of an internal complaint mechanism, which enables users to complain against content moderation decisions to the company electronically and free of charge;
- that member states establish an out-of-court dispute settlement mechanism over content moderation decisions, which means that users are entitled to select any out-of-court dispute settlement body to resolve disputes relating to content moderation decisions, including complaints that could not be resolved by means of the internal complaint-handling system; and
- relatively detailed transparency reporting on the content moderation social media companies engage in; for example, on the number of orders received from member states' authorities to act against illegal content or the content moderation engaged in on the social media companies' own initiative, including the number and type of measures taken that affect the availability, visibility, and accessibility of published content.

The Digital Services Act could, to some extent, serve as a model for regulatory proposals around the world. A clear example is the proposed Law on Freedom, Responsibility, and

Transparency being discussed in Brazil at the time of writing. Before this, the German Network Enforcement Act (NetzDG), which requires platforms to remove 'manifestly unlawful' content within 24 hours of notification under the threat of heavy fines, had already inspired several states, including Kenya, Malaysia, and India, to tighten intermediary liability laws.[36]

Discussions on how to approach platform regulation in a manner that respects human rights and how to address human rights opportunities and risks emanating from digital technologies more broadly are also taking place in international forums. For example, efforts are underway at the UN level to agree on a Global Digital Compact, addressing a number of issues including 'the promotion of a trustworthy Internet by introducing accountability criteria for discrimination and misleading content'. UNESCO is currently drafting global guidelines for regulating internet platforms, with the aim 'to inform regulatory processes under development or review for digital platforms, in a manner that is consistent with international human rights standards'.

News organisations and content moderation

The protection of media pluralism and diversity in the digital ecosystem has been of some concern for regulators. News organisations depend on platforms for access to their audiences, advertising revenue, and funding, leading to significant platform influence over their editorial, organisational, and business choices. More specifically, there has been concern about the influence of platforms' content moderation systems on the visibility, monetisation, and reach of content and accounts, and their impact on journalists' editorial independence and the financial sustainability of journalism.

The shortcomings of automated content moderation systems – discussed in more detail in the next section of this handbook – have further complicated matters for media actors. Because of their inability to properly consider context, automated tools are also prone to removing public interest reporting, including, for example, reporting on extremist groups or human rights violations.

Example

A media outlet from Sarajevo was prevented from publishing content on Facebook that reported on the judgment by the International Criminal Tribunal for Yugoslavia (ICTY) in the case of Ratko Mladić, who was convicted of several international crimes, some in connection with the Srebrenica massacre. This was because Facebook had falsely labelled the ICTY as a criminal organisation. The automated content moderation systems therefore concluded that the article on Mladić's trial was seeking to promote a 'criminal organisation'.[37]

Some proposed regulatory frameworks include provisions that would shield 'editorial content' from content moderation rules (while other regulatory proposals, such as those in Australia or Canada, have focused more on compelling social media companies to pay for the news they use). For example, the UK's Online Safety Bill – under negotiation at the time of writing – exempts news publisher content from its content moderation rules, excluding it from the scope of the obligations that the Bill imposes on social media companies. A similar media exemption was discussed during the negotiations on the EU Digital Services Act but was ultimately rejected. The European Media Freedom Act, also under negotiation at the time of writing, proposes a process through which media service providers could request special treatment from platforms when it comes to the way their content is moderated.[38] More specifically, the current proposal provides that before any content was suspended, social media companies would need to communicate the reason for their decision to the media outlet and guarantee that any complaints by the provider 'will be processed and decided upon with priority and without undue delay'.[39]

The question of whether it is appropriate to create special content moderation rules for news media providers is contested. Some argue that social media companies' terms of service should not prevail over media organisations' own editorial standards and that media actors are typically accountable through laws, ethical rules, and membership in professional associations.

However, freedom of expression and digital rights organisations, including ARTICLE 19, have maintained that despite the undisputed challenges faced by media actors in navigating platforms' content moderation rules – and the fact that the power asymmetry between social media companies and media actors requires regulatory solutions – it is problematic from a freedom of expression perspective to afford special treatment for certain media actors when it comes to the moderation of their content. As a matter of principle, the speech of some actors should not be valued more than that of others by virtue of who they are rather than what they say.

Considering freedom of expression standards on 'hate speech' in particular, as explained earlier, the position and influence of the speaker is one of the key factors that courts have to consider when assessing whether speech reaches the level of incitement to hatred that is prohibited under Article 20(2) of the ICCPR. Asking social media companies to remove hateful content posted by ordinary users but to protect the same content if published by exempt media actors would therefore run counter to these standards.

Carve-outs for media are also likely to reinforce the power of incumbents at the expense of citizen journalists, smaller bloggers, or activists who do not fulfil the criteria for regulatory exemptions, including when they engage in journalistic activity for non-profit purposes. Freedom of expression organisations therefore generally reject any kind of media privilege when it comes to content moderation.

Content moderation processes

Social media companies employ a range of approaches to content moderation, and they use a varied set of tools to enforce content policies and restrict or remove illegal or objectionable content and accounts. Given the volume of content that users produce – some of the latest figures show, for instance, that every single minute at least 350,000 tweets are posted on Twitter, 500 hours of video are uploaded to YouTube, and more than 510,000 comments and 136,000 photos are posted on Facebook – most of the largest social media companies have opted to rely to a large extent on automated tools to reduce

the need for time-consuming human moderation, reserving the latter for specific instances of content review.

The shortcomings of automation

Automated content moderation involves the use of automated detection, filtering, and moderation tools to flag, separate, and remove particular pieces of content or accounts. There are a host of automated tools, many fuelled by artificial intelligence and machine-learning, that can be deployed during the content moderation process. These tools can be deployed across a range of categories of content and media formats at different stages of the content lifecycle to identify, sort, and remove content. Some of the most widely used automated tools and methods include digital hash technology,[40] image recognition,[41] or natural language processing (NLP).[42]

Platforms will often apply so-called hybrid content moderation, which typically involves using automated tools to flag and prioritise specific content cases for human reviewers who make a final judgement.[43]

It is only possible for platforms to moderate content at scale if they rely to some extent on automated content moderation, since human moderation would be unable to process the amount of information generated by users. At the same time, these tools can have serious risks from a freedom of expression perspective, in particular when applied to complex categories of speech such as 'hate speech' or 'terrorist' content. Such categories require a level of enhanced contextual understanding and nuance that the tools are unable to provide.

Lack of accuracy and reliability

The accuracy of a content moderation tool in detecting and removing content is highly dependent on the type of content it is trained to tackle. For instance, automated tools can be effective in identifying content that qualifies as child sexual abuse material. In this case, there is a clear international consensus that the content is illegal, there are clear

parameters for what should be flagged, and models have been trained on enough data to yield high levels of accuracy.[44]

This is not the case when it comes to categories such as 'hate speech' or 'terrorist content'. In order for NLP classifiers to be trained to operate accurately, they need to be given clear parameters and definitions of speech.[45] But definitions of extremist or terrorist content – let alone what exactly may constitute praise or support for a terrorist organisation – are notoriously nonexistent or vague. More generally, NLP tools are often unable to comprehend the nuances and contextual elements of speech or to identify when content is satire or published for reporting purposes.[46] It is essential that journalists and human rights organisations are able to raise awareness about terrorist atrocities: screening and removing 'terrorist content' without contextual appreciation risks interrupting legitimate journalistic coverage and documentation of human rights violations. There are various instances where automated moderation has resulted in overbroad takedown of public interest content.

Example

The Syrian Archive, a project that aims to preserve evidence of human rights violations and other crimes committed during the conflict in Syria for the purposes of advocacy, justice, and accountability, has found that videos documenting war crimes were removed from YouTube. This can lead to widespread and sometimes permanent loss of what might be crucial evidence of war crimes.

Example

In May 2021, the American Civil Liberties Union (ACLU) reported that dozens of Tunisian, Syrian, and Palestinian activists and journalists covering human rights abuses and civilian airstrikes complained that Facebook had deactivated their accounts pursuant to its policy against 'praise, support, or representation of' 'terrorist groups', pointing to the combined issues of 'blunt automated detection systems' and the lack of an internationally agreed definition for terms like 'terrorism', 'violent extremism', or 'extremism', let alone 'support' for them.

The accuracy of NLP classifiers particularly decreases when applied across different languages and contexts. Indeed, automated tools are limited in their ability to parse and understand variances in language and behaviours that may result from different demographic and regional factors.[47] In addition, most NLP tools have lower accuracy when parsing non-English text due to a lack of resources in other languages.[48]

Amplification of bias

Beyond the issue of diminished accuracy when analysing non-English languages, the presence of bias in automated tools runs the risk of further marginalising and censoring groups that already face disproportionate prejudice and discrimination online and offline. This risk originates in the many types of human biases that are fed into training data and therefore can be amplified through the use of automated tools.[49] If training datasets are not sufficiently representative, there is a risk that artificial intelligence (AI) systems learn and perpetuate any underlying bias in the data.[50]

Example

The ACLU has reported that for years, Meta treated the speech of women and people of colour differently than that of men and white people – including when describing their experiences of sexual harassment and racism. In 2017, when women of colour and white people posted the exact same content, Meta suspended only the accounts of women of colour.

Lack of transparency and accountability

Greater transparency and accountability are needed around the application of automated tools. There is a lack of transparency around how datasets are compiled, how accurate automated content moderation tools are, and how much content is removed, both correctly and incorrectly.[51] This raises concerns about the freedom of expression rights of individuals whose content has been mistakenly flagged or whose accounts have been erroneously removed. These concerns are further exacerbated if upload filters are used and the flagged content disappears from the platforms before it is even posted, making it challenging to even know if the content has been removed in error.[52]

User reporting and 'trusted flaggers'

Typically, social media companies allow their users to report content which they believe to be illegal, in breach of their community standards, or simply harmful. As mentioned earlier, some regulatory systems attach legal consequences, in terms of intermediary liability, to such reporting through 'notice and takedown' procedures, which is problematic from a freedom of expression perspective and increases the risk of vexatious or abusive notices.

Some social media companies also rely on a 'trusted-flagger' system, whereby reports filed by trusted flaggers are fast-tracked for review. Typically, trusted flaggers are

either individuals or entities with specific expertise to identify and flag illegal content. They may include civil society actors.

Social media companies currently generally provide insufficient information about the trusted-flagger system. This includes how trusted flaggers are selected and the extent to which content flagged by trusted flaggers is subject to adequate review or is automatically removed. Although the trusted-flagger system may contribute to better-quality notices, it is not equivalent to an impartial or independent assessment of the content at issue. Trusted flaggers are often identified due to their expertise on the impacts of certain types of content, whether copyright, terrorism-related content, or 'hate speech', and their proximity to victims of such speech, but not on the basis of having freedom of expression expertise. They are therefore not necessarily well placed to make impartial assessments of whether restricting the content at issue is consistent with international human rights law.[53]

The EU Digital Services Act has incorporated this system, giving special privilege to trusted flaggers of content and stating that once notified by trusted flaggers, platforms must remove illegal content 'expeditiously'. Civil society has criticised the Act for allowing governmental and law enforcement agencies to be awarded the status of trusted flaggers, which could open the door to abuse.

The need for deeper understanding of context

The shortcomings of automated content moderation systems make it indispensable that social media companies employ sufficient human reviewers and invest resources to better understand the context of particular forms of expression. This is necessary in order to review and adequately assess categories of speech that are not amenable to being assessed by automated moderation, and to better address content moderation decisions that users appeal against. In particular, human reviewers need to be properly vetted and trained to reduce the risk of bias in their decision-making. They should also be native in the languages they cover and have sufficient knowledge of the local contexts applicable. There is currently a significant lack of transparency

as to how many human content moderators the respective social media companies employ, how companies allocate moderation tasks per country and language, how they are trained, the specific issues they respond to, and where they are located. This lack of transparency points to a broader issue of insufficient resources allocated by the largest social media companies to understanding the content in many countries where they operate but which they do not consider of strategic importance.

Conclusion

Content moderation is a fast-moving field. Regulatory environments keep changing, new actors emerge, social media companies update their policies regularly, and international organisations keep launching initiatives to find global or regional solutions to 'problematic' online content. Content moderation also needs to constantly respond to real-world challenges, like the Covid-19 pandemic, international and non-international armed conflicts, and new products and technologies such as the metaverse.

Despite these rapid changes, certain aspects of content moderation will not be subject to change any time soon. Human rights must lie at the heart of any attempt by international organisations or states to regulate social media companies. Social media companies must become serious about living up to their responsibility to respect human rights, including when it comes to content moderation, and to understand the context in which they operate.

So far, they have failed to do so. The lack of investment in deep understanding of local contexts includes a failure to properly engage with local civil society actors to acquire that understanding. Local coalitions on freedom of expression and content moderation can bridge the gap between local stakeholders and social media companies and ensure that human rights and relevant local contexts are appropriately integrated into content moderation decisions. But this requires serious commitment from social media companies to respect human rights and to mitigate adverse human rights impacts that result from shortcomings in their content moderation systems and processes.

ARTICLE 19's report *Content Moderation and Freedom of Expression* elaborates these issues in more detail and summarises what is at stake: 'When these companies fail to take into consideration the various (linguistic, political, social, cultural, and economic) dimensions of local contexts, content moderation processes can have dramatic impacts on the societies affected, such as increasing polarisation and the risk of violence.'

Bibliography

ARTICLE 19, *Hate Speech Explained: A Toolkit*, 2015.

ARTICLE 19, *Online Harassment and Abuse against Women Journalists and Major Social Media Platforms*, 2020.

ARTICLE 19, *Side-Stepping Rights: Regulating Speech by Contract. Policy Brief*, 2018.

ARTICLE 19, *Watching the Watchmen: Content Moderation, Governance, and Freedom of Expression. Policy Brief*, 2021.

UNESCO, *Addressing Hate Speech on Social Media: Contemporary Challenges*, 2021.

UNESCO, 'Countering Hate Speech'.

UNESCO, *Finding the Funds for Journalism to Thrive: Policy Options to Support Media Viability*, 2022.

UNESCO, 'How to Address Online #HateSpeech with a Human-Rights Based Approach?', 2022.

UNESCO, 'The Rabat Plan of Action on the Prohibition of Incitement to Hatred', YouTube video, 2022.

UNESCO, *Safeguarding Freedom of Expression and Access to Information: Guidelines for a Multistakeholder Approach in the Context of Regulating Digital Platforms*, 2023.

UNESCO, 'Towards Guidelines for Regulating Digital Platforms for Information as a Public Good', YouTube video, 2023.

UNESCO, *Windhoek+30 Declaration: Information as a Public Good*, World Press Freedom Day International Conference, November 2021.

Endnotes

[1] The terms 'hate speech' and 'disinformation' are not defined in international human rights law. For these reasons, ARTICLE 19 uses these terms in inverted commas throughout this publication.

[2] Through its adoption in a resolution of the UN General Assembly, the UDHR is not strictly binding on states. However, many of its provisions are regarded as having acquired legal force as customary international law since its adoption in 1948; see *Filartiga v. Pena-Irala*, 630 F. 2d 876 (1980) (US Circuit Court of Appeals, 2nd circuit).

[3] UN General Assembly, International Covenant on Civil and Political Rights, 16 December 1966, UN Treaty Series, vol. 999, p. 171.

[4] Article 10 of the European Convention for the Protection of Human Rights and Fundamental Freedoms, 4 September 1950; Article 9 of the African Charter on Human and Peoples' Rights (Banjul Charter), 27 June 1981; Article 13 of the American Convention on Human Rights, 22 November 1969.

[5] See European Court of Human Rights, *Dink v. Turkey*, paras. 106 and 137 (Applications no. 2668/07, 6102/08, 30079/08, 7072/09, and 7124/09), 14 September 2010.

[6] A. Kuczerawy (2017) 'The Power of Positive Thinking, Intermediary Liability and the Effective Enjoyment of the Right to Freedom of Expression', JIPITEC, p. 226.

[7] UN Human Rights Committee, General Comment No. 34 on Article 19: Freedoms of Opinion and Expression, CCPR/C/GC/34, 12 September 2011, paras. 12, 17, and 39.

[8] Article 10 of the European Convention for the Protection of Human Rights and Fundamental Freedoms; Article 9 of the Banjul Charter, 27 June 1981; Article 13 of the American Convention on Human Rights, 22 November 1969. For an explainer of the Three-Part Test, see UNESCO, 'The Legitimate Limits to Freedom of Expression: The Three-Part Test', YouTube video.

[9] ARTICLE 19's 'Hate Speech' toolkit provides a guide to identifying 'hate speech' and how to effectively counter it while protecting the rights to freedom of expression and equality. See also UNESCO, *Addressing Hate Speech on Social Media: Contemporary Challenges*, and UNESCO, 'How to Address Online #HateSpeech with a Human Rights-Based Approach?', YouTube video.

[10] See also the UNESCO video, 'The Rabat Plan of Action on the Prohibition of Incitement to Hatred'.

[11] Throughout this report, we refer to and quote from the community guidelines, terms of service, and other similar documents of various social media platforms, as they were at the time of writing. These are living documents that are often changed or updated, and they may no longer contain the material being discussed.

[12] For more on the different types of false information, see UNESCO, *Journalism, Fake News & Disinformation*.

[13] The Council of Europe report *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making* advocates the following distinction between 'misinformation', 'disinformation', and 'malinformation': misinformation is when false information is shared, but no harm is meant; disinformation is

when false information is knowingly shared to cause harm; malinformation is when genuine information is shared to cause harm, often by moving information designed to stay private into the public sphere.

[14] Recent examples include Turkey, Tunisia, Sudan, and the UK.

[15] In this context it is important to note that creating an enabling environment for the rights to freedom of expression and equality that addresses the underlying causes of both 'disinformation' and 'hate speech' is essential. States should, for example, focus on their positive obligations to promote a free, independent, and diverse communications environment, including media diversity and digital and media literacy, which are key means of addressing 'disinformation' and 'hate speech'. For further details, see UNESCO, 'Countering Hate Speech'; ARTICLE 19, *'Hate Speech' Explained: A Toolkit*; ARTICLE 19, 'Submission to UN Special Rapporteur on Freedom of Expression and "Disinformation"'; and UNESCO, *Media and Information Literate Citizens: Think Critically, Click Wisely!*

[16] *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*, developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises. The UN Human Rights Council endorsed the *Guiding Principles* in its resolution 17/4 of 16 June 2011.

[17] *UN Guiding Principles*, Principle 15.

[18] UN Human Rights Council, Report of the Special Rapporteur on Freedom of Expression, 6 April 2018, A/HRC/38/35, paras. 45–46.

[19] UN Doc., A/HRC/38/35, para. 52.

[20] UN Doc., A/HRC/38/35, paras. 28 (for proportionality of sanctions) and 59 (for effectiveness of remedies).

[21] UN Doc., A/HRC/38/35, para. 11.

[22] UN Doc., A/HRC/38/35, para. 50.

[23] UN Doc., A/HRC/38/35, para. 51.

[24] UN Doc., A/HRC/38/35, para. 52.

[25] See ARTICLE 19, *Watching the Watchmen*, p. 16.

[26] ARTICLE 19, *Content Moderation and Freedom of Expression*, p. 18.

[27] For best practices, see Spandana Singh and Kevin Bankston, The Transparency Reporting Toolkit: Content Takedown Reporting.

[28] ARTICLE 19, Side-Stepping Rights, pp. 16-17.

[29] ARTICLE 19, *Internet Intermediaries: Dilemma of Liability*, p. 3.

[30] The Court of First Instance found that, as administrator, it was Chiranuch Premchaiporn's responsibility to monitor the content on the forum closely, because it could negatively affect the national security and rights and freedoms of others. The decision was upheld by the higher courts.

[31] For a summary of the decision, see Columbia University Global Freedom of Expression, *Prosecutor v. Chiranuch Premchaiporn*.

[32] Special Rapporteur on the Protection and Promotion of Freedom of Opinion and Expression, Report of 16 May 2011, A/HRC/17/27, para. 43.

[33] UN Doc., A/HRC/17/27, para. 42.

[34] See, for example, the UK's Online Safety Bill or the EU's Digital Services Act.

[35] See, for example, the UK Online Safety Bill (although the 'legal but harmful' for adults provision has since been replaced by an obligation for companies to enforce their Terms of Service) or the French Draft Bill on Countering Online Hatred (so call Loi Avia or Avia Bill), which the French Constitutional Council (Conseil d'État) subsequently declared unconstitutional.

[36] J. Mchangama and J. Fiss, *The Digital Berlin Wall: How Germany (Accidentally) Created A Prototype for Global Online Censorship*; see additional comments by the authors in 'The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two'.

[37] ARTICLE 19, *Content Moderation and Local Stakeholders in Bosnia and Herzegovina*, p. 39.

[38] Article 17 of the proposed European Media Freedom Act.

[39] Article 17, proposed European Media Freedom Act.

[40] Hash matching assigns a unique digital 'fingerprint' to previously detected harmful images and videos. Newly identified harmful user-generated content can then be automatically removed if the computed hash matches a hash stored in the database of known harmful content. See Ofcom, *Use of AI in Online Content Moderation*, p. 48.

[41] While digital hash technologies use image recognition, the technique can also be used more broadly – for instance, to identify specific objects within an image, such as a weapon. See S. Singh (2019) 'Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content', *New America*, July, p. 14.

[42] NLP is a technique by which text is parsed in order to make predictions about meaning – for example, whether it expresses a positive or negative opinion. See Center for Democracy and Technology, 'Mixed Messages? The Limits of Automated Social Media Content Analysis', p. 9.

[43] Singh, 'Everything in Moderation', p. 7.

[44] Singh, 'Everything in Moderation', p. 7.

[45] Center for Democracy and Technology, 'Mixed Messages?', p. 5.

[46] Singh, 'Everything in Moderation', p. 13.

[47] Singh, 'Everything in Moderation', p. 18.

[48] ARTICLE 19, *Content Moderation*, p. 17; Brennan Center for Justice, *Double Standards in Social Media Content Moderation*, p. 18.

[49] Center for Democracy and Technology, 'Mixed Messages?' p. 6.

[50] Center for Democracy and Technology, 'Mixed Messages?' p. 14.

[51] Singh, 'Everything in Moderation', p. 16.

[52] B. Heller (2019) 'Combating Terrorist-Related Content through AI and Information Sharing', Institute for Information Law, 26 April, p. 3.

[53] ARTICLE 19, *Side-Stepping Rights*, p. 32.