July 2025
Deputy-Chair
*Leiden Model United Nations 2025*

# Balancing content moderation on social media with freedom of expression

*Economic and Social Council (ECOSOC)*

# Table of Contents


**Introduction** **2**

**Defenition of key terms** **3**

**General overvieuw** **5**

**Major parties involved** **7**

**Timeline of events**
**9**

**Possible solutions** **10**

**Further reading** **11**

**Bibliography**
**12**

# Introduction

Social media platforms such as Instagram, X (formerly Twitter), TikTok, and Facebook allow billions of people to express themselves, share opinions, and access information. However, these same platforms also create space for the spread of harmful or illegal content, including hate speech, misinformation, and incitement to violence.

To address these risks, platforms engage in content moderation. This involves deleting posts, banning users, or labelling content that could cause harm. Moderation is essential for protecting users, but if it becomes too strict, it can limit freedom of expression, a fundamental human right that protects the ability to express ideas, even when controversial or unpopular.

The challenge is finding the right balance. If platforms moderate too little, harmful content can circulate freely. This could lead to real-world harm such as violence, discrimination, or manipulation of public opinion. If moderation is too strict, it risks censorship, particularly of minority groups, activists and opposing opinions.

Different governments take different approaches to this issue. The European Union's Digital Services Act requires large platforms to quickly remove harmful content and publish transparency reports. The United Kingdom's Online Safety Act focuses on protecting users, especially minors, though some argue it could threaten privacy and free speech. In the United States, the First Amendment strongly protects freedom of speech, yet recent political debates over alleged government pressure on platforms have raised concerns about indirect censorship.

In short, social media must be both a safe environment and a place where people can speak freely. Finding the right balance between protecting users and preserving freedom of expression lies at the heart of this global challenge.

# Definition of key terms

### *Algorithmic bias*

When a platform's automated system (algorithm) treats certain types of content or users unfairly, often without meaning to. For example, it might promote some opinions more than others or hide certain voices based on language, race, or political views.

### *Censorship*

When information is blocked or removed to stop people from seeing or sharing it. This can be done by governments, companies, or platforms when content is considered harmful, illegal, or too sensitive.

### *Community guidelines*

The rules set by a social media platform explaining what users are allowed to post. These usually forbid hate speech, harassment, violent content, and other harmful posts.

### *Content moderation*

The process where platforms check what users post and decide whether to allow it, hide it, or remove it. This helps to stop the spread of harmful or illegal content.

### *Deplatforming*

When a person or group is completely removed from a platform for breaking rules. This includes deleting their account or blocking their content permanently.

### *Disinformation*

False information shared on purpose to trick or mislead people. It is often used to influence opinions, spread fear, or cause confusion.

### *Freedom of expression*

The right of people to share their opinions and ideas freely in any form, as long as they do not break the law or harm others. This right is protected in international law.

### *Hate speech*

Words or messages that attack or insult people based on their race, religion, gender, or other personal traits. Hate speech can lead to harm, violence, or discrimination.

### *Misinformation*

False or wrong information shared by mistake, not to trick people. This can happen when people believe something is true without checking the facts.

### *Online platform*

A website or app where people can post and share content with others. Examples include TikTok, YouTube, Instagram, X (formerly Twitter), and Reddit.

### *Shadow banning*

When a platform secretly hides a user's posts so fewer people see them, without telling the user. The account still works, but its content becomes less visible.

### *Terms of service*

The legal contract that users accept when joining a platform. It describes the rules they must follow and what the platform can do with their content.

### *Transparency reporting*

When platforms publish reports explaining how they moderate content. These reports usually include the number of posts removed, users banned, and requests from governments.

# General Overview

Social media platforms such as TikTok, Instagram, Facebook, and X (formerly Twitter) play a central role in modern communication. These platforms allow billions of people to share ideas, express opinions, and access information. However, this open environment also allows harmful or illegal content to spread, including hate speech, misinformation, and incitement to violence.

To address these risks, platforms use content moderation: actions like removing posts, banning accounts, limiting the reach of certain content, or adding warning labels. Human moderators or automated systems (algorithms) can take these actions. The primary goal is to prevent harm while maintaining the platforms' openness to free expression.

Balancing safety and freedom of expression is a significant challenge. Overly strict moderation can limit freedom of speech, potentially silencing activists, minorities, or critical opinions. Inadequate moderation may allow harmful content to proliferate unchecked, posing a risk of social unrest, misinformation, or public harm.

### *Types of harmful content*

**Hate speech**: Content that attacks individuals or groups based on attributes such as race, religion, gender, or nationality.

**Misinformation and disinformation**: False or misleading content that can confuse or mislead the public.

**Incitement to violence**: Content that encourages or promotes acts of violence or illegal actions.

Different platforms define these categories in their policies and apply moderation based on their guidelines and national laws.

### *Approaches by governments*

Governments around the world control content moderation in different ways, depending on their legal systems, cultural values, and national interests. Some prioritise freedom of expression, whilst others prioritise safety, national security, or cultural values.

The European Union (EU)
In 2023, the EU implemented the Digital Services Act (DSA). This law requires huge platforms to promptly delete illegal information and explain how their algorithms work. Companies must provide reports on how they manage content moderation. If companies fail to meet these requirements, they may risk fines of up to 6% of their global revenue. The DSA aims to make social media safer while preserving freedom of expression.

United Kingdom (UK)
In 2023, the UK passed the Online Safety Act. This regulation aims to protect people, especially minors, from dangerous information like cyberbullying and online abuse. Platforms must take measures to stop the spread of dangerous or unlawful content. However, some experts believe that the law may limit the freedom of speech because the definition of damaging content is not always clear. Ofcom, the UK's media regulator, oversees the law's enforcement.

United States (US)
In the United States, the First Amendment gives strong protection to freedom of speech. The government cannot directly order companies to remove content except in extreme cases like threats of violence. However, there has been growing debate about the government's informal influence over social media platforms. For example, during the COVID-19 epidemic and the 2020 election, government agencies were accused of pressing companies to remove particular posts. Some of these cases are currently under review by US courts.

Other countries
Some countries, such as China, Russia, and Iran, strictly control online content. Governments in these countries block platforms or require companies to remove political content. These policies are often explained as ways to protect national security or cultural values. However, human rights organisations have criticised them for limiting freedom of expression.

Other countries, including Australia and Canada, are developing laws that focus on transparency and user protection. They aim to regulate harmful content without restricting free speech.

### *Role of private companies*

The platforms where content moderation occurs are run by private tech firms. These companies are in charge of hiring human moderators, using automated tools, and creating content restrictions. Some businesses stress how difficult it is to moderate a lot of content in a variety of languages and cultural contexts.

At the same time, civil society groups and some governments have raised concerns about transparency and accountability in moderation practices. Questions have been raised about how algorithms operate, how decisions are made, and whether content removal affects certain groups more than others.

Many companies have introduced transparency reports and independent oversight boards to explain their moderation processes and decisions. Stakeholders continue to evaluate the effectiveness of these strategies.

### *International perspectives*

Different countries and organisations have different opinions on where to draw the line between harmful content and free speech. Some prioritise public safety and the welfare of society, while others focus on protecting personal freedoms and reducing government interference. International cooperation is difficult because there is currently not a universal agreement on content moderation standards.

### *Real-world impacts*

The way content moderation is applied can have significant consequences. Insufficient moderation may lead to the spread of harmful content, while overly strict moderation may risk limiting legitimate speech. Social media platforms, governments, civil society, and users continue to discuss the best way to manage these risks.

# Major parties involved

### *European Union (EU)*

The European Union is one of the most active authorities in terms of online content control. With the adoption of the Digital Services Act (DSA), the EU imposed tough requirements on large platforms, forcing them to rapidly delete illegal content and make transparency reports detailing how their moderation procedures work. The EU's approach focuses on making online places safer while protecting freedom of expression under EU law. Historically, the EU has prioritized regulating the digital market to protect consumers, particularly privacy (GDPR) and disinformation. Its policies are intended to apply to all member states.

### *Meta Platforms, Inc. (Meta)*

Meta owns platforms including as Facebook, Instagram, and Threads, making it a key player in global discussions about content moderation. The company detects and removes harmful content using artificial intelligence, human moderators, and automated tools. Meta frequently adjusts its moderation procedures to meet local restrictions, which means that policies may differ across nations. Meta usually supports clear regulations, but has been criticized for inconsistent enforcement and its involvement in promoting misinformation during elections or emergencies. The company also participates in global industry alliances to combat hazardous content, such as terrorism and child exploitation.

### *TikTok / ByteDance*

TikTok is a global social media platform owned by the Chinese company ByteDance. It uses both automated systems and human moderators to manage content. However, the platform often faces criticism for being unclear about how it decides what content to remove or promote. In several countries, TikTok has been banned or restricted due to concerns over data privacy, national security, and the potential for content manipulation. Critics worry that because ByteDance is based in China, the company might be influenced by Chinese government policies, possibly leading to hidden censorship. ByteDance denies these claims, stating that it follows local laws and applies its moderation rules fairly. TikTok's approach to moderation differs from that of Western companies, reflecting different risks and priorities.

### *United Nations Educational, Scientific and Cultural Organization (UNESCO)*

UNESCO is an international organization that promotes freedom of expression both online and offline. It works to guide governments and companies in creating digital policies that protect human rights while ensuring user safety. In 2023, UNESCO published principles for regulating digital platforms, warning against over-censorship and highlighting the need for transparency in how content is moderated. The organization encourages governments to protect free speech and avoid silencing dissent, while also addressing harmful content like hate speech and misinformation. UNESCO also offers educational resources and support to help countries develop fair and balanced online regulations.

### *United States of America (USA)*

In the United States, freedom of speech is strongly protected by the First Amendment, which limits how much the government can control online content. As a result, content moderation has mostly been handled by private companies. Recently, however, this system has been questioned. There have been accusations that the US government has indirectly influenced platforms to remove certain content, especially during elections or public health emergencies. This has led to public debates and legal cases over the risk of hidden censorship. At the same time, some politicians are pushing to change Section 230 of the Communications Decency Act, a law that shields platforms from being held legally responsible for user content. Proposed reforms aim to make tech companies more accountable for harmful content without undermining free speech protections.

# Timeline of events

| | |
|---|---|
| **2004 Febuary 4th** | Facebook is founded, marking the rise of large-scale social media platforms allowing widespread public content sharing. |
| **2016 November 8th** | Twitter begins implementing more aggressive conent moderation policies during global events like the London Olympics, highlighgting early large-scale moderation efforts. |
| **2018 May 25th** | The General Data Protection Regulation (GDPR) comes into force in the European Union, setting data privacy standards that indirectly affect content management. |
| **2020 March 11th** | The World Health Organization declares COVID-19 a global pandemic, leading to new policies by platforms to curb health misinformation. |
| **2022 October 27th** | The European Union officially adopts the Digital Services Act (DSA), mandating faster removal of illegal content and requiring transparency from large online platforms. |
| **2023 February 7th** | The United Kingdom passes the Online Safety Act, imposing duties of care on online platforms to tackle harmful content and protect young users. |
| **2023 November 9th** | UNESCO publishes global Guidelines for Regulating Digital Platforms, encouraging governments and companies to balance safety with freedom of expression. |

# Possible solutions

### *More Transparency from Social Media Platforms*

Social media companies should regularly publish clear reports showing how they decide which posts to remove or block. These reports can explain how many posts are taken down, for what reasons, and who makes those decisions. This would help users understand how content moderation works and reduce the risk of hidden censorship.

### *Independent Review Panels*

Platforms can set up independent groups of experts who check difficult content moderation cases. These panels would help make sure decisions are fair and balanced. Governments or organizations like UNESCO could help set the rules for how these panels should work.

### *Better Laws Across Countries*

Countries could work together to create similar rules for content moderation. Right now, every country has different laws, which makes things confusing for platforms and users. International agreements could help protect freedom of speech while keeping people safe online, no matter where they are.

### *Stronger Appeal Options for Users*

Users should have a simple way to challenge a content moderation decision if they think it is unfair. Platforms should explain clearly why content was removed and allow users to appeal quickly and easily. This helps protect freedom of expression and builds trust between users and platforms.

### *Control Over Algorithms*

Platforms should be more open about how their algorithms work, especially those that decide which content people see. Governments could require platforms to make sure their algorithms do not promote harmful content by accident. This would help reduce the spread of dangerous information without deleting large amounts of content.

# Further reading


- ARTICLE 19. (2025). Digital rights resources and advocacy. https://www.article19.org/issues/digital/
  Focused on protecting digital freedom and fighting censorship.

- OHCHR. (2025). Freedom of expression and media guidelines. https://www.ohchr.org/en/topics/freedom-expression
  UN reports and legal frameworks addressing online freedom of speech.

- UNESCO. (2023). Regulating digital platforms without harming freedom of expression. https://www.unesco.org/en/digital-platforms
  Practical guidance for balanced content moderation strategies.

- Center for Democracy and Technology (CDT). (2024). Content moderation toolkit. https://cdt.org/issue/content-moderation/
  Resources on how companies and policymakers can design fair moderation systems.

- Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
  Academic analysis of moderation practices.

- Internet Governance Forum (IGF). (2024). Dynamic Coalition on Platform Responsibility. https://www.intgovforum.org/en/
  International dialogues on regulating tech platforms.

# Bibliography


ARTICLE 19. (2025, May 22). Protecting digital freedom in a conflicted world: Priorities for RightsCon 2025. https://www.article19.org/resources/rightscon-2025-protecting-digital-freedom-conflicted-world/

BBC News. (2024, May 5). Free speech vs safety: Global battle over social media laws. https://www.bbc.com/news/technology

European Commission. (2024). The Digital Services Act (DSA). European Union. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act

Freedom House. (2024). Freedom on the Net 2024: The repressive power of artificial intelligence. https://freedomhouse.org/report/freedom-net/2024/repressive-power-artificial-intelligence

Meta Platforms Inc. (2024). Content governance and transparency reports. Meta Transparency Center. https://transparency.fb.com/

OHCHR. (2025). Reports and statements on freedom of expression and social media. United Nations Office of the High Commissioner for Human Rights. https://www.ohchr.org/en/search?query=social+media+freedom

Reuters. (2024, March 18). TikTok faces bans, regulatory scrutiny as governments question moderation policies. https://www.reuters.com/technology

TikTok. (2024). Community guidelines enforcement report Q1 2024. https://www.tiktok.com/transparency/en/

Twitter Inc. (2024). X Transparency Center: Content moderation updates. https://transparency.twitter.com/en/reports

UK Government. (2023). Online Safety Act 2023: Summary and guidance. https://www.gov.uk/government/collections/online-safety-bill

UNESCO. (2023). Guidelines for regulating digital platforms: Safeguarding freedom of expression and access to information. United Nations Educational, Scientific and Cultural Organization. https://www.unesco.org/en/digital-platforms

U.S. Congress. (2024). Section 230 of the Communications Decency Act. https://www.congress.gov/