unesco

UNITED NATIONS OFFICE ON GENOCIDE PREVENTION AND THE RESPONSIBILITY TO PROTECT

Addressing hate speech on social media: Contemporary challenges

Discussion paper

01

This paper was produced by researchers at the Oxford Internet Institute, with the support of UNESCO, as a contribution to the United Nations Strategy and Plan of Action on Hate Speech, as well as within the framework of the project "#CoronavirusFacts: Addressing the 'Disinfodemic' on COVID-19 in conflict-prone environments" funded by the European Union. It is part of the Oxford Internet Institute's collaboration with UNESCO to develop a toolkit that maps existing methods, resources, and research projects developed to monitor the existence, spread and impact of online hate speech, as well as assess capacities and practices to counter it. Any comments in relation to this discussion paper that could help to inform the larger study are welcomed.

Addressing and countering hate speech is a multilayered endeavour, which includes tackling its root causes and drivers, preventing it from translating into violence and dealing with its wider societal consequences. To develop effective responses to hate speech, including through education, it is essential to better monitor and analyse the phenomenon by drawing on clear and reliable data. In the digital age, this also means better understanding the occurrence, virulence and reach of online hate speech.

The identification of online hate speech for research purposes is confronted with numerous challenges, from a methodological perspective – including definitions used to frame the issue, social and historical contexts, linguistic subtleties, the variety of online communities and forms of online hate speech (type of language, images, etc.). From a technological perspective, online hate speech is difficult to study due to the inconsistent reliability of detection systems, opaque nature of proprietary algorithms, lack of access to data held by companies and so forth. Clarity on how these challenges can be addressed is indispensable for producing further understanding of how online hate speech emerges and proliferates, and subsequently for formulating effective responses.

The United Nations Strategy and Plan of Action identifies a series of priority areas for monitoring and analysing hate speech, stipulating that relevant United Nations entities should be able to "recognize, monitor, collect data [on] and analyse hate speech trends". When focusing on online hate speech, United Nations entities are encouraged to promote "more research on the relationship between the misuse of the Internet and social media for spreading hate speech and the factors that drive individuals towards violence", as well as "map the emerging risks and opportunities related to the spread of hate speech posed by new technologies and digital platforms"; and finally, "to define action protocols that account for the new forms of digital hate speech".

Over the past year, the coronavirus disease (COVID-19) pandemic has further underlined the pertinence of the United Nations Strategy and Plan of Action on Hate Speech, as a wave of hate speech spread across the world, further exacerbating intolerance and discrimination towards particular groups and destabilizing societies and political systems.

This discussion paper seeks to give an overview of the key aspects that need to be taken into consideration to address the occurrence of hate speech on social media, be it through concrete regulations by social media companies, counter efforts and legislations or preventive educational measures. The paper is divided into three sections: part 1 focuses on definitions of hate speech and associated legal frameworks, part 2 reviews and addresses tools and techniques for monitoring hate speech online and discusses measurements of the prevalence of online hate speech and part 3 discusses potential counter and preventive measures.

1
DEFINING HATE SPEECH

The difficulties of addressing and legislating about hate speech begin with its definition. There is no internationally accepted definition of hate speech. Instead, it presents several legal issues such as freedom of opinion and expression, discrimination and the advocacy or incitement of discrimination, hostility or violence.

The Dangerous Speech Project by Susan Benesch[1] suggests that there are two main difficulties with the term 'hate speech'. Firstly, 'hate' is a vague term that can take on different levels of intensity and be followed by different consequences – "does the "hate" in hate speech mean that the speaker hates, or seeks to persuade others to hate, or wants to make people feel hated?"[2] Secondly, 'hate speech' at its core means that people or a group are targeted because of their identity/membership of a group. This requires that a law or definition has to specify whether or not it considers all identities and groups to fall under this law, and if not, what kind of groups are included. The Dangerous Speech Project argues that overly broad laws can be misused against vulnerable groups or political and civic opposition, at times resulting in harming the same groups that hate speech laws aim to protect. However, it could also be argued that a definition that too narrowly focused on specific groups and identities could lead to legal exclusion or lack of legal tools to address the problem.

While the scope of the present discussion paper does not allow to examine these challenges in detail, a review of international and national laws around the world illustrates the complexities and different interpretations that can be applied to hate speech.

At the global level, alongside the non-binding *Universal Declaration of Human Rights*, the *International Covenant on Civil and Political Rights* (ICCPR) follows up the right to freedom of expression (Article 19) with a prohibition of any advocacy of hatred that constitutes incitement to discrimination, hostility or violence (Article 20). Articles 19 and 20 also place limitations on restricting freedom of expression – these can "only be such as are **provided by law** and are necessary: **(a) for respect of the rights or reputations of others; (b) for the protection of national security or of public order, or of public health or morals**."

Complementing these principles, the *Rabat Plan of Action* proposes a "**six part threshold test**" to justify restrictions on freedom of expression, considering the **socio political context, status of the speaker, intent to incite antagonism, speech content, extent of dissemination and likelihood of harm**.

Also prominent on hate speech is the *International Convention on the Elimination of All Forms of Racial Discrimination* (ICERD), which outlines a stricter clause than the ICCPR's Article 20, as it does not require intent or the 'advocacy of hatred' and includes dissemination in the list of punishable practices. Further relevant to this space are, among others, the *Convention on the Prevention and Punishment of the Crime of Genocide* and the *Convention on the Elimination of All Forms of Discrimination against Women* (CEDAW).

The freedom of expression organization ARTICLE 19 has developed the *Camden Principles on Freedom of Expression and Equality* on the basis of discussions with United Nations officials and experts from academia and civil society. These principles provide interpretive guidance to ICCPR articles and seek to deter actors from abusing Article 20 by specifying issues around 'incitement', as well as what constitutes 'discrimination', 'hostility' and 'violence'.

[1] The Dangerous Speech Project. 2021. https://dangerousspeech.org/
[2] Benesch, S. 2021. Dangerous Speech: A Practical Guide. The Dangerous Speech Project, 2021, p. 7. https://dangerousspeech.org/

When translating international law and principles into national law, each country has a slightly different approach to how it defines hate speech in terms of how it is expressed, who the potential targets are and what kind of harm has to happen for speech itself to be considered hateful. The lack of a unified definition is one of the major challenges when it comes to combatting online hate speech, which is not necessarily confined to national borders.

Definitions of hate speech also matter for research and advocacy efforts, particularly when identifying its societal consequences. Harms resulting from hate speech may be located at the level of individuals (in the form of psychological harm), groups and communities, and society (in the form of the erosion of rights and public goods). Since hate speech targets people on the grounds of group characteristics, analysis at the level of community harm is particularly important. Harms due to hate speech are unequally distributed across the general population, with marginalized groups bearing the brunt of the burden. They are also cumulative for those who suffer them, with prior experience of hate speech being a key variable in estimating the harm derived from being targeted with hate speech.[3]

ONLINE HATE SPEECH

Hate speech online is not intrinsically different from hate speech offline. However, it differs in the nature of the interactions in which it takes place/occurs, as well as in the use and spread of specific words, accusations and conspiracy theories that can evolve, peak and fade very quickly. Hateful messages can go viral in hours or even minutes.

The 2015 UNESCO report *Countering Online Hate Speech* points out how online hate speech can be produced and spread at low cost, does not go through an editing process like other written work, experiences vastly different levels of exposure depending on the popularity of the post and can be posted cross-nationally, as platform servers and headquarters do not need to be in the same country as the user and their intended audience. Hate speech online can also be available for longer and go through waves of popularity, connect with new networks or reappear, as well as be anonymous.

Consequently, the question of who moderates online spaces and if and when content should be removed has been widely debated.

This debate is exemplified by laws such as the **German Network Enforcement Act (NetzDG)**, which was presented in 2017 and requires social media platforms with more than 2 million users to implement a transparent procedure to moderate illegal content (which includes hate speech), remove content that has been identified as illegal within a timeframe of 24 hours and regularly report on the measures taken. The law was heavily criticized as pushing platforms into a role of "privatized censorship" for decisions that should be made by courts and it was warned that the time limits and the fines would lead to platforms "over-removing" content to avoid the risk of high penalties. In 2020, the law was revised to require social media platforms to forward identified illegal content to the Federal Criminal Police Office. Another, almost simultaneous, revision strengthened users' rights by requiring platforms to make the reporting of illegal content more user-friendly and to enable the appeal of a decision to delete or not delete a post.

In this vein, the development of laws to address online and offline hate speech are often fraught with complicated review processes related to both definitional challenges and the task of respecting freedom of expression within the framework of the law. Given these challenges, methods going beyond legal measures must also be employed to address hate speech.

[3] Gelber, K. and McNamara, L. 2016. Evidencing the harms of hate speech. Social Identities, Vol. 22, No. 3, pp. 324 341.

2
TOOLS AND TECHNIQUES FOR MEASURING AND MONITORING HATE SPEECH

Policies and tools regarding the detection, monitoring and moderation of online hate speech vary across contexts, actors and platforms.

Detection methods can be broadly grouped into two categories: more comprehensive efforts that have initially relied on keyword filters and crowd-sourcing methods, and those that rely on human content moderators who review content that has been flagged as hate speech by users and decide on whether it classifies. Whilst manual approaches have the distinct advantage of capturing context and reacting rapidly to new developments, the process is labour-intensive, time-consuming and expensive, limiting scalability and rapid solutions. As a result of these challenges, the increasing volume of content produced on social media and the advances in machine learning and natural language processing techniques, platforms and researchers have developed and increasingly rely on automated detection solutions. Many newer initiatives use a variety of methods in combination. Some key terms related to these methodologies include:

- **Machine learning:** Techniques that utilize computer algorithms that can improve automatically through experience and by the use of data.
- **Natural language processing:** Techniques that process and analyse large amounts of natural language data.
- **Keyword-based approaches:** Methods using an ontology or dictionary, identifying text that contains potentially hateful keywords.
- **Distributional semantics:** Methods for quantifying and categorizing similarities between words, phrases and sentences based on how they are distributed in large samples of data.
- **Sentiment analysis:** Methods to explain what kind of attitudes are conveyed in relation to a subject in a given text.
- **Source metadata:** Some methods inform models through the meta-information of the data, such as data about the users associated with the messages, including network based features such as their number of followers.
- **Deep learning:** A class of machine learning algorithms that use multiple layers to progressively extract higher-level features from the raw input.

Social media companies have largely shifted from reacting to posts flagged by users as hate speech to proactively detecting and addressing such content through their automated systems, before users have seen it. While necessary to address hate speech at scale, these methods also introduce complications: automated hate speech detection inevitably makes mistakes and the removal of non-hateful content is possible. Excessive content removal could create chilling effects and undermine free speech.

To improve their monitoring activities, hate speech detection tools are continuously being developed. **Perspective API** [4], for instance, is an open-source tool by Jigsaw (an incubator within Google) and Google's Counter Abuse Technology team and has been used by news organizations and Google's products. It uses machine learning to score phrases based on their potential toxicity in a conversation. It is available in seven

[4] More information about Perspective API is available at https://www.perspectiveapi.com/

languages (English, French, German, Italian, Portuguese, Spanish and Russian). **Facebook** has stated that the latest version of its tool to monitor and detect hate speech on its platforms has improved in its semantic understanding of language and understanding of content, including its analysis of images, comments and other elements [5].

Researchers and civil society organizations have also worked to develop tools for detecting hate speech. Some examples include:

- The Kenyan platform **Umati** was one of the first to manually monitor online posts written in relevant languages in Kenya.
- Davidson et al. (2017) developed **HateSonar**, using a logistic regression approach trained on data from Web fora and Twitter.
- The Anti Defamation League (ADL) and UC Berkeley's D-Lab developed the **Online Hate Index** (OHI), designed to transform the human understanding of hate speech via machine learning into a scalable tool that can be deployed on internet content to discover the scope and spread of online hate speech.
- The Measures & Counter-measures project team at the Alan Turing Institute developed a tool using deep learning methods **to detect East Asian prejudice** on social media.
- Moon et al. (2020) developed a **Korean Hate Speech Detection** tool. They trained the model with both "bias" and "hate" labels.
- **Hatemeter** detects Anti-Muslim hate speech using machine learning and natural language processing techniques. The platform is available in English, French and Italian.
- **COSMOS** collects and analyses data from Twitter in real time by keyword specification, using sentiment analysis and natural language processing.
- **MANDOLA** detects hateful content through a combination of sentiment analysis, natural language processing, machine learning and deep learning.

It is important to note that monitoring online hate speech is dependent on access to data, especially from social media platforms. Furthermore, today, many of the existing tools are monolingual and often limited to the English language and more research is needed on the performance of multilingual detection methods. Additionally, the vast majority of the research on and monitoring of hate speech on social media platforms has focused on the United States and Europe, leading to a gap not only in tools and data, but in the understanding of the extent and dynamics of the spread of hate speech in other regions. This gap is all the more crucial to bridge given the contextual nature of hate speech.

PREVALENCE OF HATE SPEECH ON SOCIAL MEDIA PLATFORMS

Using automated detection tools based on methods available today, Twitter, Facebook, Instagram and YouTube have increasingly reported flagged and/or removed content. Between January and March 2021, YouTube removed 85,247 videos that violated their hate speech policy. Its two previous reports show similar figures. For the same quarter, Facebook reported a total of 25.2 million pieces of content actioned, whilst Instagram reported 6.3 million pieces of content. According to Twitter's last transparency report, the company removed 1,628,281 pieces of content deemed to violate their hate speech policy between July and December 2020.

On social media platforms, the prevalence of hate speech is determined through a sample of content that users view. In other words, it captures only (an estimate of) what hate speech remains on the platform beyond what the company has already proactively detected and removed. To date, Facebook is the only platform that reports prevalence metrics. The company reported that between January 2021 and March 2021, there was a 0.05% to 0.06% prevalence of hate speech, showing a slight decrease compared to their two previous reports. Some studies indicate that the prevalence of hate speech on mainstream platforms such as Twitter and Wikipedia is less than 1% of total content, whilst on more niche alternative platforms such as Gab and 4chan, between 5% and 8% of content hosted may be abusive [6]. Evidence about the prevalence of hate speech on social media platforms remains incomplete, partly due to a lack of transparency and data access on the part of platforms.

[5] Source: https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/
[6] Zannettou et al. 2018; Mathew et al. 2018; Hine et al. 2017; Vidgen et al. 2019.

3
COUNTERING HATE SPEECH ONLINE

It is first worth highlighting that countering hate speech – and, by extension, selecting appropriate tools and strategies as well as preventive efforts – is complicated by several factors. There is little consensus in different actors' answers to key questions in a range of contexts. How does hate speech harm, and when is harm severe enough to warrant the regulation of speech? More granularly, which kinds of harms, linked to which hateful speech behaviours, warrant regulations in line with international human rights law and freedom of expression standards?

The architecture of the internet also adds distinct challenges to countering hate speech. These include the permanence, itinerancy, anonymity and cross-jurisdictional character of online content, the wide range of platform architectures and a heterogenous, multi-stakeholder system of internet governance.

Despite these challenges, many groups and individuals are engaged in various ways of combatting hate speech online and preventatively strengthening the resilience of online users to it.

STATE-LEGAL RECOURSE

A salient avenue for countering hate speech is legal recourse. Although stances on hate speech and hate speech online vary between regions and continue to evolve as the issue is better understood, there exists a number of international principles, regional agreements, state-level laws and examples of jurisprudence in alignment with international human rights standards that have clauses relevant to online and offline hate speech, as outlined in the beginning of this paper.

However, issues with strictly legal responses to hate speech online quickly arise. These include concerns with the balance of rights, the possibility of powerful actors abusing the restrictions of rights and a reliance on thresholds for the prohibition of incitement to violence alongside a poorly understood relationship between hate speech and violence offline. More importantly for the countering of online hate speech, a key issue for legal recourse is the limitation of individual States' authority over online digital spaces. Effectively addressing online hate speech cannot solely rely on national legal recourse.[7]

In 2016, a group of major tech companies agreed upon the European Commission's *Code of Conduct on Countering Illegal Hate Speech Online*, which requires these companies to review hateful speech within a day of receiving a report. This approach is challenging due to high variance in terms of service and operational definitions of hate speech, but it is a significant effort in promoting collaboration and linking legal and extra-legal approaches in the hate speech space.

RESPONSES BY TECH COMPANIES

In 2021, both YouTube and Facebook reported an increase in the content found and flagged by each respective company, along with a greater proportion of content flagged by the company compared to content flagged by users. This is due to a growing use of automated detection systems. However, the trend is accompanied by an increase in restored content in relation to the previous reported periods. Between January and March 2021, Facebook restored 408,700 pieces of content and Instagram 43,700. Whilst the reports suggest that platforms are increasingly actioning hateful content, what we do not know is whether this is due to an accompanying increase in the amount of abuse, an increase in the stringency of the platforms' policies or an increase in false positives.

Social media companies are based in national jurisdictions, therefore directly impacted by national laws and usually more responsive to requests to contain hate speech as a result. Social media platforms, however,

[7] Since 2013, UNESCO's Judges Initiative has raised the capacities of judicial actors on international and regional standards on freedom of expression, access to information and the safety of journalists in regions across the world, particularly motivated by the fact that the question of how best to treat hate speech cases is one of the key interests of many judicial operators. Over 23,000 judicial actors have been trained on these issues, notably through a series of massive open online courses (MOOCs), on-the-ground training and workshops, as well as the publication of a number of toolkits and guidelines.

are not bound by territoriality and are therefore relied upon only to uphold their own terms of service, which may or may not be stricter than the standards set out by the international agreements outlined in a previous section. Actions taken by social media platforms include removing material when it is judged to be hate speech, as well as sending warnings to users posting hateful speech, restricting their activity on the platform or banning them. These community standards are constantly evolving, particularly in how much they rely on automated versus human moderation methods.

In light of these challenges, a multi-stakeholder movement calling for greater transparency of internet companies as a means of enhancing their accountability has gained growing momentum in recent years. This has included proposed legal and regulatory measures in at least 30 countries and regions, including through the European Digital Services Act currently in development. Companies have also taken steps to be more transparent. In 2021, Access Now indexed over 70 companies that issue regular transparency reports [8], although much more is needed.

The UNESCO issue brief titled *Letting the Sun Shine In: Transparency and Accountability in the Digital Age* presents enhancing transparency as a third way between state overregulation of content, which has led to disproportionate restrictions on human rights, and a laissez-faire approach that has failed to effectively address problematic content such as hate speech and disinformation. The brief provides a set of 26 high-level principles spanning issues related to content and process, due diligence and redress, empowerment, commercial dimensions, personal data gathering and use, and data access.

EXTRA-LEGAL RESPONSES AND PREVENTIVE INTERVENTIONS

Other extra-legal responses come from the research and advocacy efforts of civil society or may focus on preventive measures that strengthen the resilience of online users to hate speech. These include initiatives that directly target the causes and consequences of online hate speech, including through education, as well as initiatives that call for better legal and tech based measures to be implemented.

Education-based initiatives are at the heart of these efforts and often focus on long-term prevention. Educational interventions can serve to raise awareness about the harmful consequences of hate speech, address its root causes and effectively alert to manipulation techniques and rhetoric used to spread hate, online and offline. In particular, media and information literacy programmes have been developed and implemented around the world, aiming to provide online users with the skills to critically examine online content and identify disturbing, hateful content and misinformation. Similarly, counterspeech efforts aimed to address hate speech with positive counternarratives such as the Facebook-led Online Civil Courage Initiative, conducted in Germany, the United Kingdom of Great Britain and Northern Ireland and France in 2017, have also been made. Other civil society initiatives focus on advocating for the platforms to bring about change. In July 2020, the Stop Hate for Profit campaign brought together a coalition of over 1,200 companies from around the world and called for an ad boycott against major platforms, demanding hate speech moderation and an ad pause on accounts that promote discrimination against certain groups. This campaign added to a growing number of voices that called for addressing online hate speech, particularly in light of the intensified targeting of marginalized groups during the COVID-19 pandemic, prompting several social media companies to make changes to their community guidelines.

[8] Access Now. https://www.accessnow.org/transparency-reporting-index/

RECOMMENDATIONS

To inform evidence-based policymaking to curb online hate speech - and to prevent hate speech from translating into violence while also safeguarding freedom of expression - it is critical to recognize, monitor, collect data on and analyse hate speech trends in order to identify appropriate strategies to address them. The recommendations below aim to identify key actions to tackle new challenges in emerging viral hate speech and particularly to address their offline consequences for peace, stability and the enjoyment of human rights for all.

1. Promote inclusive definitions of hate speech that respect freedom of expression

- Ensure that definitions are in line with international standards, particularly as stipulated by the ICCPR and the Rabat Plan of Action.

2. Build multi-stakeholder coalitions

- Encourage the sharing of data and expertise between human rights organizations, internet intermediaries and the public.
- Empower stakeholders and notably local communities to monitor and detect hate speech on social media tailored to their context and languages.
- Convene multi-stakeholder dialogues on hate speech trends, occurrence and how to counter it.
- Advocate for platforms to develop definitions and operational routines in collaboration with expert groups and the public, which should reach outside North America and Western Europe to include more countries around the world.

3. Gather data and encourage open data practices whereby data is already collected, while respecting personal data protection

- Gather qualitative data with individuals targeted by hate speech to better understand the scope and nature of harms.
- Advocate that internet platform companies improve their transparency practices, including by openly releasing data about hate speech complaints and their resolution, as well as about the accuracy and functioning of their content moderation systems, particularly for research purposes.
- Support the development of affordable, accessible and user-friendly tools and methodologies that can be used to monitor and detect hate speech across multilingual, multicultural contexts within a timeframe that allows for counteraction.

4. Encourage platforms to offer robust remedial options for those whose content has been removed

- Facilitate collaboration between social media companies and civil society groups focused on digital rights to ensure that content moderation and removal processes are aligned with community needs.

5. Develop media and information literacy and digital skills via education programmes

- Provide funding and resources for the development of educational programmes that foster resilience to hate speech, informed by current hate speech trends and responding to related challenges. This requires a close collaboration between social media companies, research institutes and education stakeholders.
- Prioritize preventive educational approaches that alert to the harmful effects of online hate speech and foster media and information literacy alongside mitigation and counter efforts.
- Establish and support partnerships between educational institutions and social media companies to increase access to information and resources to address hate speech on social media platforms via targeted dissemination campaigns or the redirection of users to external resources.

6. Support active organizations in the online hate speech space

- Ensure that adequate resources are provided for specialized organizations dedicated to monitoring and countering hate speech, particularly those best equipped to take local contexts into account, and provide them with support.

This paper is part of a collection of discussion papers, commissioned and produced by UNESCO and the United Nations Office of the Special Adviser on the Prevention of Genocide (OSAPG).The papers are a direct contribution to the United Nations Strategy and Plan of Action and are published in the context of the Multistakeholder Forum and Ministerial Conference on Addressing Hate Speech through Education in September and October 2021.

The outbreak of the COVID-19 pandemic has highlighted the pertinence of the United Nations Strategy and Plan of Action, generating a wave of hate speech across the world –further exacerbating intolerance and discrimination towards particular groups and destabilizing societies and political systems. The discussion papers seek to unpack key issues related to this global challenge and propose possible responses and recommendations.

This paper was commissioned by the UNESCO Section of Freedom of Expression and Safety of Journalists as part of the project "#CoronavirusFacts, Addressing the 'Disinfodemic' on COVID-19 in conflict-prone environments" funded by the European Union. It was drafted by Jonathan Bright, Antonella Perini, Anne Ploin and Reja Wyss of the Oxford Internet Institute at the University of Oxford.

Published in 2021 by the United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France

This document was produced with the financial support of the European Union. Its contents are the sole responsibility of the authors and do not necessarily reflect the views of the European Union.

Graphic design: Dean Dorat

CI/FEJ/2021/DP/01

Funded by the European Union

Educate Against #HateSpeech