Organization for Security and Co-operation in Europe
Office of the Representative on Freedom of the Media

Expert workshop on hate speech

14 April 2021

CONCEPT NOTE

Due to a high quantity and low quality of user-generated content shared on large online platforms, effective content moderation is increasingly necessary – and at the same time, difficult to achieve. In order to scale content moderation, private actors adopt automated content detection tools to grapple with societal phenomena such as hate speech. These tools, however, fail to assess the contextual nature of information as well as nuances of human communication due to their current contextual blindness. Moreover, in recent years, the prevailing risk of discriminatory biases against marginalised groups that are inherent in these systems have been largely documented by a number of human rights organisations around the world.[1]

The goal of this expert group is to identify the actual and foreseeable negative impact of automated tools for detecting and evaluating online hate speech imposed on individuals' human rights, with the emphasis on the right to freedom of expression and opinion of marginalised groups, and media freedom as such, and to develop recommendations to effectively address this impact.

The expert group will not seek to determine what constitutes online hate speech, as there is no universally adopted definition of hate speech at international level. From a content moderation point of view, it can be argued that this is the most challenging category of user-generated content to be identified and detected. Moreover, there are forms of hate speech expressions that fall into the category of user-generated content, which is potentially harmful but legal. However, even if hate speech expression falls into the protective realm of the right to freedom of expression and opinion, it may still have discriminatory impact, carry potential collective harm and silence marginalised groups.[2] This raises the question of how to address and tackle the impact of potentially harmful but legal speech, while ensuring full respect for the right to freedom of expression.

Especially for potentially harmful but legal expressions, it is highly relevant how these categories of user-generated content are being defined by the terms of service formulated and enforced by

[1] Cobbe, J. (2020). Algorithmic Censorship by Social Platforms: Power and Resistance', Philosophy & Technology, Philosophy & Technology.
[2] Llanso E, Hoboken J, Leersen P, et al. (2020). Artificial intelligence, content moderation, and freedom of expression. Trans-Atlantic Working Group Working Papers Series.

online platforms. Due to their dominance and power over the public sphere, internet intermediaries are capable of setting the standard for what is permitted online globally. Moreover, they are developing and deploying the technologies used to implement this standard, to the detriment of transparency and accountability.[3]

First, the expert group will pay particular attention to discriminatory bias imposed by automated tools. The work will focus on two main ways how such discriminatory bias manifests in online space in relation to the right to freedom of expression and opinion:

- The ability to safely participate in online platforms is critical for marginalised groups to form a community and find support.[4] Automated tools develop their ability to identify and distinguish different categories of content based on the datasets they are trained on. If these datasets do not include examples of speech in different languages and from different groups or communities, they will not be equipped to parse these groups' communication. Automated tools may either miss the potentially hateful content by generating false negatives or wrongfully label legitimate expressions as hate speech, so-called false positives. This way, those targeted by online hate speech remain without any effective remedy against abuse; while at the same time, other, legitimate speech may be unjustifiably restricted.

- The impact of discriminatory bias can manifest as "biased censorship" against content posted by groups and their members who are at the same time often targeted by hateful expressions and online abuse. While hate speech itself is difficult to automatically identify and remove, groups likely to be targeted by online abuse and hate speech may themselves find their communications censored and thus, being silenced. Applying a tool to a domain or group of speakers who do not closely match the groups represented in the training data can lead to erroneous classifications that disproportionately affect marginalised groups. Hence, automated tools developed with the purpose to identify "toxic speech" can themselves introduce further collective harm by failing to recognise the context in which speech occurs and thus, reinforcing harmful stereotypes against marginalized groups.[5] Therefore, the expert group should provide recommendations to identify, analyse and assess significant systemic risks stemming from content

[3] The example of the latest Facebook hate speech policy demonstrates this alarming trend quite well. For further details, please consult Access Now, Why Facebook's proposed hate speech policy on Zionism would only add fuel to the fire, 2021.

[4] Tomasev N., McKee, K., Kay J., Mohamed S. (2021). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities.

[5] A recent study demonstrated that an existing toxicity detection system would routinely consider drag queens to be as offensive as white supremacists in their online presence. The system further specifically associated high levels of toxicity with words like 'gay', 'queer' and 'lesbian'. For further details, please consult Gomes A., Antonialli D., Dias Oliva T., Drag queens and Artificial Intelligence: should computers decide what is 'toxic' on the internet?, 2019.

moderation systems against marginalised groups and their negative impact on their participation in public discourse.

Second, the goal of the expert group is to provide recommendations on strengthening the position of those targeted by online hate speech. The group should provide operational recommendations that will strengthen the access to effective remedy and redress, especially in cases of opaque automated decision-making processes that do not contain clear explanations and often generate unsatisfactory outcomes. Recommendations should be directed at promoting meaningful transparency and accountability.

The expert group should focus on algorithmic commercial content moderation at scale defined as "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geoblocking, account takedown).[6] The work of the expert group should explore automated systems that make decisions about content and accounts, including natural language processing (NLP), e.g. Google/Jigsaw's Perspective API. The expert group should provide guidance on independent auditing of algorithmic content moderation tools as well as (ex-ante) human rights impact assessments, with the emphasis on the need to protect those targeted by online hate speech against discriminatory biases.

In this vein, members of the expert group will provide a set of human rights centric recommendations addressed to OSCE participating States with the aim to identify effective ways to adhere to human rights obligations, due diligence standards, and procedural fairness safeguards that can effectively prevent these risks.

[6] Gorwa R, Reuben Binns R, Katzenbach C, 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance.

Organization for Security and Co-operation in Europe
Office of the Representative on Freedom of the Media

Expert Workshop on Hate Speech
14 April 2021

AGENDA

10:00 - 10:10
Welcome by OSCE RFoM and Access Now

- Teresa Ribeiro, OSCE Representative on Freedom of the Media
  - Welcoming Remarks
- Eliska Pirkova, Europe Policy Analyst, Access Now
  - Introducing the agenda and objectives of the Working group
  - Housekeeping rules

10:10 - 11:00
Tour de table

- Name and affiliation
- What potential negative impact of automated decision-making do you see, concerning measures against online hate speech, on an individual's human rights, in particular freedom of expression of marginalised groups, and the wider societal risk they can potentially carry for freedom of the media?

11:00 - 11:10
Coffee break

11:10 - 12:00
Session 1: Positive obligation of States to protect those targeted by hate speech against free expression violations committed directly or indirectly by private actors

What measures and obligations should be put in place to achieve meaningful decisional transparency and effective accountability in automated decision-making systems deployed by platforms to moderate potentially legal but harmful content, with a specific focus on hate speech?

- Introduction by the Chair covering the main areas for this session

- Discussion among experts
    - What obligations should be established by regulatory frameworks in order to enable meaningful transparency regarding the specific functionalities of automated decision-making systems aimed at combating hate speech?
    - What AI-based measures that are less intrusive for freedom of expression than removals can provide some level of protection against hate speech to marginalised groups?
    - How can States ensure that those targeted by online hate speech, including members of marginalised groups, can reach effective remedy and redress?

12:00 - 13:00
Lunch break

13:00 - 13:50
Session 2: Human rights obligations of States in respect of algorithmic systems that may negatively impact individuals' human rights, and in particular the right freedom of expression

Many automated decision-making systems combating hate speech are based on developing classifiers to categorize user-generated content. They require significant amounts of hand-labeled inputs. However, the processes of generating training datasets and having one or more people label it can introduce discriminatory biases and errors into the model. How can human rights compliance be guaranteed and algorithmic fairness be achieved in these systems?

- Introduction by the Chair covering the main areas for this session
- Discussion among experts
    - How can States ensure that marginalised groups play an active role in algorithmic design and development of automated tools moderating content, with a view to preventing the risk of impeding their right to freedom of expression?
    - How can States secure an inclusive and participatory approach when compiling training datasets used by automated tools tackling hate speech in order to prevent any potential discriminatory biases?
    - How can States secure the proper testing and evaluation of datasets used by automated content moderation as well as the functioning of the algorithmic

systems with due regard to potential discriminatory bias, with emphasis on the right to freedom of expression?

13:50-14:00
Coffee break

14:00-14:50

Session 3: Exploring positive outcomes delivered by automated decision-making systems in protecting marginalised groups against online hate speech

- Introduction by the Chair covering the main areas for this session
- Discussion among experts
    - What forms of automated decision-making systems, if any, have a positive impact on the protection of marginalised groups against online hate speech?
    - What role can automated decision-making systems play in mitigating collective harm caused by hate speech on marginalised groups and in researching the consequences of this phenomenon, so the research and evidence based outcome can inform future regulatory efforts?

14:50 - 15:00
Coffee break

15:00 - 15:50
Closing remarks:

- Brief discussion on areas not covered by this workshop which would need additional attention (in the scope of the specified subject matter: “hate speech”)
- Summarising the takeaways, seeking to identify operational recommendations from the “how to” discussions in the three sessions
- Explaining the next steps