IDI–Yad Vashem

Recommendations for Reducing Online Hate Speech

A Proposed Basis for Policy Guidelines for Social Media Companies and other Internet Intermediaries

Introduction

The recommendations presented below are the product of a yearlong study conducted by an international team of researchers,[1] with guidance from an international steering committee of experts[2] convened by the Israel Democracy Institute (IDI) and Yad Vashem. The process included workshops in Jerusalem (hosted by the IDI), Geneva (hosted by the Geneva Academy for International Humanitarian Law and Human Rights), and Irvine (hosted by the Center on Globalization, Law and Society of the University of California at Irvine), and the writing of three detailed research papers that offer multiple policy recommendations. Throughout the study, consultations were held by the research team with academics, policy researchers, government officials, human rights activists, industry policy officers, technology experts, and others.

The sixteen recommendations that emerged from the study and the research papers are meant to provide social media companies and other internet

[1] The research papers that provided serve the basis for the recommendations were written by Dr. Tehilla Shwartz-Altshuler and Mr. Rotem Medzini, by Prof. Karen Eltis and Dr. Ilia Siatitsa, and by Prof. Susan Benesch.

[2] The Steering Committee comprised the following experts: Prof. Tendayi Achiume (the UN Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance), Prof. Sarah Cleveland (former vice-chair of the UN Human Rights Committee), Prof. Irwin Cotler (former Minister of Justice, Canada), Prof. David Kaye (the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), Prof. Avner Shalev (chair of the Yad Vashem Directorate), Prof. Yuval Shany (former chair of the UN Human Rights Committee), and Prof. Jacques de Werra (Vice-Rector, University of Geneva).

intermediaries with a basis for policy guidelines and benchmarks and with directions for future action aimed at reducing hate speech and protecting the fundamental human rights the find themselves under assault by such speech, while ensuring freedom of expression (including the protection speech that may offend, shock, or disturbs the public) and other relevant human rights. They also provide other stakeholders that are troubled by online hate speech, including civil society, the public at large, and institutions invested with special responsibilities in this regard, such as elected governments and independent judiciaries, with tools to evaluate company policies and rules on hate speech and the manner of their application.

Recommendation 1: The Responsibility to Reduce Online Hate Speech

Social media companies and other internet intermediaries have a legal and ethical responsibility to take effective measures to reduce the dissemination of prohibited hate speech on their digital platforms and to address its consequences. This includes, where appropriate, content moderation (see Recommendation 6) and the recognition and condemnation of such speech. Measures such as content moderation have a critical relationship to basic human rights, including freedom of expression, the right to equal participation in political life, the right to personal security, and freedom from discrimination. Pursuant to internationally accepted legal standards and definitions, company policies and rules on prohibited hate speech must be transparent, open to independent review, and offer accessible remedies for violations of the applicable norms. The responsibility of social media companies and other internet intermediaries does not release other actors, including online users, group and page administrators and moderators, private and public associations, states, and international organizations from

their responsibility under domestic and international law to take effective measures to reduce online hate speech and their liability for the harm caused or facilitated by prohibited hate speech.

Recommendation 2: The Application of Relevant Legal Standards

Policies and rules aimed at reducing hate speech should conform to international human rights standards, as found in the International Covenant on Civil and Political Rights (especially articles 19 and 20) and in other international instruments, such as the Convention on the Elimination of Racial Discrimination (especially articles 4 and 5(d)(viii)), the European Convention on Human Rights, and other regional human rights conventions. They should conform to national laws, provided that such laws are compatible with international standards. The policies and rules should also be informed by broadly supported international instruments, such as the Rabat Plan of Action with the six potential indicators of criminal hate speech it identifies (context, speaker, content and form, extent and reach of the speech act, and likelihood, including imminence) and the Working Definition of Antisemitism adopted by the International Holocaust Remembrance Alliance.

Recommendation 3: **The Harm Principle**

In the determination of whether certain speech should or should not be considered prohibited hate speech subject to content moderation policies and rules, particular attention should be given to the need to effectively prevent harm to groups and individuals, including physical and psychological harm, reputational harm, and affront to their dignity, and to an evaluation of

whether such harm is likely to result from the speech, given the speaker's overall tone and intention, the methods and means of its dissemination, and the status of the persons targeted by the speech and/or of the protected group to which they belong, including patterns of tension and discrimination and violence against targeted protected groups, such as antisemitism, Islamophobia, and xenophobia. When denial of clearly established historical facts about the most serious international crimes, such as the Holocaust and other past genocides, is intended and expected to re-victimize victims and their descendants, it should be considered a harmful form of speech.

Recommendation 4: Detailed Policy Guidelines

Social media companies and other internet intermediaries should clearly define and publish detailed policy guidelines on prohibited hate speech and permitted speech, anchored in the applicable human rights standards. They should explain how they apply their policies and rules, especially how context—including social, cultural, and political diversity, the use of code words and euphemisms, criticism of hate speech and humor, and the reclamation of offensive slurs by targeted groups—is taken into account in decisions about content moderation. The detailed definitions of hate speech used by social media companies and other internet intermediaries should be formulated after consultation with outside experts who are familiar with the relevant national and international legal standards on hate speech, as well as with experts in other relevant fields, such as education, sociology, psychology, and technology.

Recommendation 5: Preventive Measures

Social media companies and other internet intermediaries should adopt proactive policies that are consistent with international human rights standards and that are designed to prevent the dissemination of prohibited hate speech before it causes different forms of harm. They should harness reliable algorithms for natural-language processing and reliable sentiment-analysis tools, whose decisions are subject to meaningful human review and challenge mechanisms, and employ their own internal trained content reviewers, with the aim of improving the identification of hate speech, curtailing the virality of prohibited harmful content, and/or allowing users to apply filters to block offensive content they do not wish to be exposed to. Social media companies and other internet intermediaries should also take steps to render their policies and rules visible and easily accessible to users, presented in a concise, transparent, and intelligible manner and written in clear and plain language, including examples of permissible and impermissible content. With the goal of discouraging users from resorting to hate speech, these proactive policies should be designed to foster understanding of the relevant policies and rules and employ culturally sensitive awareness-raising measures, which might include explaining how certain expressions or images might be perceived by affected individuals or groups.

Recommendation 6: A Diversity of Content-Moderation Techniques

To enforce hate-speech policies and rules, social media companies and other internet intermediaries should develop an array of content-moderation techniques that go beyond simply deleting content and blocking accounts. Such techniques should include nuanced measures that are adjusted to

different degrees of deviation from the policies or rules, the source of the complaint about a violation (e.g., an AI-based algorithm, law-enforcement agency, trusted community partner, other online user), and the identity of the speech-generating user (private individual, news agency, educational institution, repeat offender, etc.). These fine-tuned measures could include the flagging of content, the attachment of countervailing materials to potentially harmful content, a warning to disseminators of the consequences of violations, a request to disseminators to self-moderate or remove harmful content, and the unilateral imposition by the platform of temporary limits on dissemination. Special strategies need to be put in place to address chronic and particularly serious violations of hate-speech policies and rules, including the permanent blocking of repeat violators, the dismantling of business models which deliberately use online platforms to facilitate prohibited harmful activities, and notifications to law-enforcement agencies of serious violations that might merit attention by criminal justice authorities.

Recommendation 7: Flagging Mechanisms

Social media companies and other internet intermediaries should institute mechanisms that allow for a quick and effective response to the flagging of prohibited hate speech by algorithms or internal content reviewers, and for soliciting external notifications from community partners (such as law-enforcement agencies, civil society groups, and other users) and responding to them quickly and effectively. This should include the introduction of conspicuously placed standard user interfaces and national contact points for notifications. Companies and intermediaries should also rely on information

from trusted community partners in order to introduce temporary content-moderation measures, such as measures to curtail virality.

Recommendation 8: Notification of Complaints and Decisions

In order to facilitate quick and effective oversight at all stages of decision-making about content moderation, complainants must be sent immediate acknowledgement that their notification about prohibited hate-speech content has been received. Subsequent decisions about content moderation must be conveyed to them with an explanation of the reasons for the decision, including reference to any anticipated harm or lack thereof, and information on possibilities of challenge or appeal. Decisions to moderate content and the reasons for the decision must also be communicated to the user that published the speech deemed hateful.

Recommendation 9: Ordinary Mechanisms for Challenging Decisions

Social media companies and other internet intermediaries should develop effective and accessible mechanisms for challenging their specific decisions to moderate or not moderate speech alleged to be hateful, and for quickly and effectively resolving such challenges. Procedures for reviewing challenges to specific decisions should be introduced at the platform level, including an internal process for rapid reconsideration of specific decisions on content moderation, as well as access to a private alternative dispute resolution (ADR) process or litigation, when appropriate, for dealing with disputes about final decisions about content moderation which are not resolved internally.

Recommendation 10: Mechanisms for Examining 'Hard Cases'

Procedures should be developed for consulting with legal advisors or advisory bodies about specific decisions or the application of general policies or rules to a specific situation. "Hard cases" – cases where it is not readily apparent to company personnel responsible for content moderation decisions whether the speech in question conforms to or violates applicable policies and rules – should be promptly examined by independent experts. In addition, governments should ensure that content-moderation decisions that infringe the freedom of expression and other basic rights of individuals subject to their jurisdiction are subject to review by independent courts.

Recommendation 11: Protection of Content Moderators

Social media companies and other internet intermediaries should establish effective programs for training content moderators, with human rights education and cultural sensitization relevant to the content they review, including the considerations set forth in Recommendation 3.[3] They should also take adequate measures to mitigate trauma and other adverse consequences of excessive and prolonged exposure to hate speech, including setting limits on the working hours of content reviewers and providing them with counseling and other forms of psychological support.

[3] See, e.g., the following MOOCs: https://www.holocaustremembrance.com/news-archive/yad-vashem-online-course-antisemitism; https://www.mooc-list.com/course/le-racisme-et-lantis%C3%A9mitisme-fun.

Recommendation 12: Advisory Councils

Social media companies and other internet intermediaries should establish advisory councils to periodically evaluate their content moderation policies and rules and the manner in which they monitor and enforce these policies and rules, including the practice of designating cases as "hard cases," challenge procedures, and transparency policies. Such advisory councils should be composed predominantly of independent experts familiar with the applicable international standards, content-moderation technology, education policy, and relevant political, cultural, and other contexts. Where appropriate, advisory councils should be established not only at the international level, but also at the national (or regional) levels, so they can evaluate and suggest ways to adapt general policies and rules to local norms and cultural contexts without violating international human rights standards. To ensure transparency and accountability, the procedures and criteria for selecting the members of advisory councils, including safeguards against conflicts of interest, should be made public.

Recommendation 13: Exchange of Information and Best Practices among Companies

Social media companies and other internet intermediaries should consider establishing procedures (including joint advisory councils) for exchanging information about their content-moderation policies, rules, training methods, and challenge mechanisms, with a view to coordinating and, where appropriate, aligning their key elements to the best industry practices. They should also consider creating a common digital database of hashtags, images, phrases, and code words associated with prohibited hate speech, in different social, political, and cultural contexts, and, subject to privacy

constraints, sharing information about repeat violators of their hate speech policies.

Recommendation 14: A Global Stakeholders Forum

A global stakeholders forum, with representative of governments, social media companies and other internet intermediaries, experts in technology, law, and education, and civil society groups, should be created and convened from time to time in order to discuss, develop, and evaluate the application of international standards and procedures for reducing online hate speech.[4]

Recommendation 15: Transparency

Social media companies and other internet intermediaries should publish regular detailed reports on the application of their hate speech policies and rules, including country-specific information about specific content modifications, whether at the request of law-enforcement agencies or at their own initiative, information about external notifications, about challenges to specific content moderation decisions and their outcome, and about the training of content moderators, efforts to raise users' awareness, partnerships with civil society organizations, and other proactive measures. Reports on content-modification activities should be sufficiently detailed to allow external assessment of these practices' compliance with international human rights standards. In addition, information about the scale of public exposure

[4] The Global Network Initiative and the International Holocaust Remembrance Alliance are possible models for such a global coalition.

to harmful content prior to content moderation by the platform should be made available to the public.

Recommendation 16: Criteria for Evaluation of Policies and Rules

Advisory councils, civil society organizations, the media, and other observers may find it useful to evaluate and compare the policies and rules for hate-speech content moderation applied by different social media companies and other internet intermediaries, so as to encourage identification of best practices and to allow users to make more informed choices between different legitimate policies and to assess whether they adequately balance the need to address hate speech with respect for freedom of expression and other individual rights . The evaluating of hate-speech policies and rules could take the following factors into consideration:

(1) The definition of protected groups: Does it cover collectives other than racial, ethnic, and religious groups, such as those defined on the basis of their sex, sexual orientation, or gender identity, or on the basis of disability, and voluntary membership groups (e.g., political parties or professional associations)? Does the definition address situations of intersectional discrimination?

(2) The extent to which the classification of hate speech as such (a) is based on a closed list of banned words, phrases, symbols or images; (b) makes it possible to identify complex connections among language, images, and ideas that may render speech hateful in certain cultural, social, or political settings, and (c) considers the broader context that may legitimize (e.g., satire) or delegitimize the speech (e.g., bogus historical research at the service of racist causes);

(3) Is the element of causation incorporated in the definition of hate speech linked only to the expectation that it might lead to physical harm to the targeted persons? Or does it also consider nonphysical damage to potential victims, such as fear or feelings of marginalization, as well as indirect harm such as discrimination as a result of negative stereotypes and social attitudes against the protected group?

(4) Are broader socially undesirable impacts on the audience of the speech factored into content-moderation decisions – ranging from likelihood of violence to other breaches of the peace (e.g., possible social unrest) and to nonphysical long-term results, such as the fostering of a climate of growing hate and racism in society?

(5) Are content-moderation decisions based only on the speakers' explicit intent, or also on their implicit intent, or regardless of their intent?

(6) Are applicable content-moderation tools applied to speech disseminated on public platforms only, or also that intended for closed groups and sent as private messages?

(7) Does the response to a violation of hate-speech policies and rules entail only limiting its virality? Or are there other measures, such as a requesting users to remove or self-moderate the content they posted, unilateral content removal, or temporary or permanent blocking of the account?

It is recommended that companies conduct a periodic self-evaluation of their policies in light of these criteria and publish the results of the evaluation.