

Reducing Online Hate Speech

Recommendations for Social Media Companies and Internet Intermediaries

Algorithms Targeted
al Media Threats Ide
city ONLINE Re
are HARRASSMENT Inc
peech Anonymous
ness Moderators Inte
rbullying Hate Crimes
makers Videos Filters



Editor:
Yuval Shany



Reducing Online Hate Speech

Recommendations for Social
Media Companies and Internet
Intermediaries

Editor: Yuval Shany

Text Editor: Lenn Schramm
Series and Cover Design: Studio Tamar Bar Dayan
Typesetting: Ronit Gilad, Nadav Shtechman Polischuk

Cover illustration: Shutterstock

ISBN 978-965-519-294-0

No portion of this book may be reproduced, copied, photographed, recorded, translated, stored in a database, broadcast, or transmitted in any form or by any means, electronic, optical, mechanical, or otherwise. Commercial use in any form of the material contained in this book without the express written permission of the publisher is strictly forbidden.

Copyright © 2020 by the Israel Democracy Institute (RA) and by Yad Vashem
Printed in Israel

The Israel Democracy Institute
4 Pinsker St., P.O.B. 4702, Jerusalem 9104602, Israel
Tel: (972)-2-5300-800; Fax: (972)-2-5300-867
E-mail: orders@idi.org.il
Website: en.idi.org.il

Yad Vashem
Har Hazikaron, P.O.B. 3477, Jerusalem 9103401, Israel
Tel: (972) 2 6443400; Fax: (972) 2 6443569
Email: webmaster@yadvashem.org.il
Website: www.yadvashem.org

The views expressed in this book do not necessarily reflect those of the Israel Democracy Institute or those of Yad Vashem.

THE ISRAEL DEMOCRACY INSTITUTE

The Israel Democracy Institute (IDI) is an independent center of research and action dedicated to strengthening the foundations of Israeli democracy. IDI works to bolster the values and institutions of Israel as a Jewish and democratic state. A non-partisan think-and-do tank, the institute harnesses rigorous applied research to influence policy, legislation and public opinion. The institute partners with political leaders, policymakers, and representatives of civil society to improve the functioning of the government and its institutions, confront security threats while preserving civil liberties, and foster solidarity within Israeli society. The State of Israel recognized the positive impact of IDI's research and recommendations by bestowing upon the Institute its most prestigious award, the Israel Prize for Lifetime Achievement.

YAD VASHEM

Yad Vashem, the World Holocaust Remembrance Center, is the ultimate source for Holocaust education, documentation, commemoration and research. From the Mount of Remembrance in Jerusalem, Yad Vashem's integrated approach incorporates meaningful educational initiatives, groundbreaking research and inspirational exhibits. Its use of innovative technological platforms maximizes accessibility of the vast information in the Yad Vashem archival collections for an expanding global audience. With comprehensive websites in eight languages, Yad Vashem strives to meet the growing global demand for accurate and meaningful information about the Holocaust. In addition, Yad Vashem's active presence in social media offers unprecedented opportunities for rapidly communicating ideas, sharing relevant content, and engaging with and connecting to a broad and diverse public.

Yad Vashem is at the forefront of unceasing efforts to safeguard and share the memory of the victims and the events of the Shoah period; to document accurately one of the darkest chapters in the history of humanity; and to grapple effectively with the ongoing challenges of keeping the memory of the Holocaust relevant today and for future generations.

Table of Contents

Contributors	7
<hr/>	
Preface	
Yuval Shany	9
<hr/>	
A Proposed Basis for Policy Guidelines for Social Media Companies and Other Internet Intermediaries	
IDI–Yad Vashem Recommendations for Reducing Online Hate Speech	13
<hr/>	
Dealing with Hate Speech on Social Media: Towards an Interoperable Model for Addressing Racism and Strengthening Democratic Legitimacy	25
Rotem Medzini, Tehilla Shwartz Altshuler	
<hr/>	
Realigning the Law to Better Uphold the State’s Duty to Protect Human Rights: Towards an Interoperable Model for Addressing Racism and Strengthening Democratic Legitimacy	
Karen Eltis, Ilia Siatitsa	187
<hr/>	
Proposals for Improved Regulation of Harmful Online Content	
Susan Benesch	247
<hr/>	

Contributors

Susan Benesch

Faculty Associate of the Berkman Klein Center for Internet and Society at Harvard University. Benesch founded and directs the Dangerous Speech Project (dangerousspeech.org), which studies speech that can inspire violence, and looks for ways to prevent this without infringing freedom of expression. To that end, she conducts research on methods of diminishing harmful speech online or the harm itself.

Prof. Karen Eltis

Professor of Law at the University of Ottawa and past Affiliate of the Center for Information Technology Policy at Princeton University. Eltis specializes in cybersecurity and democratic governance in the digital age. She served as senior advisor to the National Judicial Institute (Canada) and taught at Columbia Law School. Eltis is a research associate at the Federmann Cyber Security Center at the Hebrew University of Jerusalem. Her research on privacy has been cited by the Supreme Court of Canada.

Adv. Rotem Medzini

A doctoral candidate at the Federmann School of Public Policy and Government and a research fellow at the Federmann Cyber Security Center – Cyber Law Program, both at the Hebrew University of Jerusalem. Medzini is also a former researcher at the Israel Democracy Institute. His research interests are regulatory governance, self-regulation, and regulatory intermediation theories. Rotem's research focuses on internet policy, media and content regulation, data protection law and policy.

Prof. Yuval Shany

Vice-president for research at the Israel Democracy Institute, the incumbent of the Hersch Lauterpacht Chair in Public International Law at Hebrew University, and former dean of the Hebrew University Law Faculty. Member and former chair of the UN Human Rights Committee.

**Dr. Tehilla Shwartz
Altshuler**

Senior fellow at the Israel Democracy Institute and head of its Democracy in the Information Age Program. She is a senior lecturer at the Federmann School of Public Policy and Government and a research associate at the Federmann Cyber Security Center – Cyber Law Program, Hebrew University of Jerusalem. She is also a member of the boards of the Israel National Press Council, the Israeli Digital Rights Movement, and the Israeli National Archive Committee. She is an expert on media regulation and the interface between technology, law, and policy.

Dr. Ilia Siatitsa

Program Director and Legal Officer at Privacy International. She focuses on that organization's research, advocacy, and litigation related to surveillance and technology and heads a project that challenges the use of mass surveillance for protecting civic spaces. A qualified lawyer in Greece, she holds a Ph.D. in International Law from the Faculty of Law of the University of Geneva, and specializes in international human rights law and public international law.

Preface

Yuval Shany

This publication (which was sent to print in the autumn of 2019) contains the results of a joint research project undertaken by the Israel Democracy Institute (IDI) and Yad Vashem, with the goal of supporting efforts by social media companies and other internet intermediaries to formulate policy and policy guidelines aimed at reducing online hate speech. Although hate speech is certainly not a new phenomenon, digital platforms facilitate its promulgation and dissemination today at unprecedented speed and scale, and this requires a more proactive response to its harmful consequences. The utilization of private platforms for spreading hate in digital space also poses unique governance challenges and demands new approaches to content regulation and institutional oversight.

As a nonpartisan Israeli think tank, the IDI has a longstanding interest in the possibilities and challenges that new technologies pose to traditional democratic values, processes, and institutions. It sees the digital space as a crucial asset for democratic life in the twenty-first century, but one that must be protected against abuse. Yad Vashem, too, dedicated to perpetuating the memory of the Holocaust and the lessons learned from that dark chapter in modern history, views the digital space as an important educational arena and tool. It is concerned, however, about the malicious exploitation of this platform to spread hateful propaganda, including anti-Semitic Holocaust denial. It was against this background that these two Israel-based institutions joined forces with international partners to devise and carry out a research program to address the problem of online hate speech from a broad and nonlocal perspective. It is clear that, as a matter of principle, the ways of dealing with anti-Semitic hate speech should not be developed separately from ways of countering other vile and potentially harmful forms of hate speech that promote Islamophobia, homophobia, hatred of migrants, and the like. Only on the basis of broadly accepted norms and processes that identify prohibited hate speech and restrict it can specific modalities be devised to deal with specific forms of hate speech or to protect specific groups of potential victims.

The present publication has two sections. The first of them consists of the *IDI–Yad Vashem Recommendations for Reducing Online Hate Speech*: sixteen recommendations meant to serve as the basis of policy guidelines for social media companies and other internet intermediaries. These recommendations derive from the research papers presented in the second section, as well as from discussions undertaken in the three workshops, held in Jerusalem, Geneva, and Irvine, California, as part of the research project, and consultations among the project researchers and steering committee. The studies by members of the research team—Dr. Tehilla Shwartz-Altshuler (IDI) and Mr. Rotem Medzini (IDI), Prof. Karen Eltis

(University of Ottawa) and Dr. Ilia Siatitsa (Geneva Academy of Human Rights and International Humanitarian Law), and Prof. Susan Benesch (Berkman Klein Center, Harvard University)—analyze online platforms' current policies and the legal frameworks in which they operate, and propose avenues for future reforms. We hope that the recommendations and the research papers they are based on will inform contemporary debates on the ways in which social media companies and other internet intermediaries regulate online speech and will influence the positions of the stakeholders who participated in such debates—states, international organizations, academia, civil society, the technology sector, the media, and the public at large.

I would like to thank the administrative and policy teams at the IDI and Yad Vashem that facilitated the organization and operation of the research project, and especially Ms. Shirli Ben-Tolila (IDI), Mr. Arnon Meir (IDI), Ms. Iris Rosenberg (Yad Vashem) and Dr. Robert Rozett (Yad Vashem). Thanks are also due to Mr. Dvir Kahana, the director general of the Israel Ministry of Diaspora Affairs, and Mr. Yogev Karasenty, a senior policy officer in that ministry, for their ongoing engagement with the research project and their keen interest in its findings.

Prof. Yuval Shany
Project Coordinator
Jerusalem, 2019

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 experts² 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.

1 The research papers that form 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).

The sixteen recommendations that emerged from the study and the research papers are meant to provide social media companies and other internet 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 that find themselves under assault by such speech, while ensuring freedom of expression (including the protection of speech that may offend, shock, or disturb 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 (e.g., 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, be 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, and 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 steps 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. These 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 on 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 under 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.

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

³ See, e.g., the following MOOCs: Yad Vashem Online Course on Antisemitism, November 11, 2018; Le racism et l'antisémitisme (FUN).

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) level, 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 the formation of 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 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 representatives 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 of 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 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, to allow users to make more informed choices between different legitimate policies, and to enable users to

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

assess whether they adequately balance the need to address hate speech with respect for freedom of expression and other individual rights. The evaluation 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 in 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 request that users 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.

Dealing with Hate Speech on Social Media

Towards an Interoperable Model for Addressing Racism and Strengthening Democratic Legitimacy

Rotem Medzini | Tehilla Shwartz Altshuler

Abstract | Introduction: Hate Speech on Social Media | Chapter 1. The Legal Framework: International and National Interpretations of Hate Speech | Chapter 2. The Different Categories of Players and the Responsibility of Internet Intermediaries | Chapter 3. A Typology of Legal and Regulatory Instruments for Moderating Hate Speech on Social Media | Chapter 4. The Proposal: A Co-regulation Model with Common Criteria for Defining Hate Speech | Chapter 4(a). Common Criteria Definition for Hate Speech | Chapter 4(b). Procedures for Identifying Common Criteria and Content Moderation | Chapter 5. Advantages and Disadvantages of the Proposed Co-Regulatory Model | Appendix A. Defining the Hate-Speech Policy Problem | Appendix B. Examples of Content Moderation by Several Major OSPs | Appendix C. Applying the Proposed Model to Twitter's Community Standards

ABSTRACT

Though the need to prevent hate speech was not born with social-media platforms, the increase in the volume of hate speech on social media has negative social implications. This development demands a public

discussion about defending the right to free speech and the need for policy tools to deal with hate speech.

The proposed model provides scales and guidance to help online platforms define their preferred policy for combating hate speech. The model is co-regulatory and has two key aspects: (1) five common criteria for identifying hate speech, and (2) a detailed procedure for their application. Our criteria identify factors that categorize speech as hate speech or as speech that might lead to hate-related offenses. These factors are associated with what most countries and most major platforms would define as hate speech.

Our analysis of the criteria builds on the idea of creating a continuum of scalable options for each criterion. Using these criteria, the management of online service providers (OSPs) can decide how to implement each criterion and whether it should be implemented in a lenient or stricter manner.

(1) Does the speech target a group or an individual as a member of a group? The most basic criterion for recognizing hate speech is that the utterance targets a group or targets an individual as a member of a group. This criterion distinguishes “hate speech” from other forms of harmful speech, such as defamation, bullying, and various personal threats. Management has to decide whether it should protect only the most conservative definitions of protected groups, or whether their policies also protect other groups that people are part of – voluntary or not.

(2) Does the speech express hatred? Our continuum aims at identifying the mere existence of hate speech (rather than how extreme it is). Here the continuum starts with a closed list of banned expressions and symbols and ends with a more context-based approach that examines content in its context. In the middle are policies that build on natural language processing to mimic how human content moderators label problematic content.

(3) Could the speech cause harm to an individual or a group? This criterion asks whether the content aims to cause additional harm beyond the speech itself. Here the continuum ranges from physical harm to non-physical and indirect mental harm.

(4) Does the speaker intend to harm? The importance of intent as a factor, despite the difficulties of identifying it, derives from its close connection to the ability to cause actual harm. Here policies can range from searching for explicit intent, to using human or natural language processing capabilities in order to identify implicit consent, to ignoring the speaker's intent altogether.

(5) Does the speech incite to socially undesirable actions?

Our model also includes a co-regulatory implementation mechanism in which OSPs and law-enforcement agencies share responsibility for moderating hate speech: OSPs devise the procedures and implement measures, and law-enforcement agencies notify them of problematic content. Our model, however, does not challenge the OSPs' current upload practices or deal with their policies regarding page and group managers.

The first step of our co-regulatory execution mechanism is the implementation of the common criteria described above, as a function of where an OSP's decision-makers choose to locate its policy on the various scales. The type of speech is also a factor to be considered, because different policy rules may apply to public statements than to open groups, closed groups, or private messages. Based on their financial and technological abilities, OSPs should develop algorithm-based instruments for active monitoring and automatic flagging of questionable content, train human content moderators, and diversify their staff to reduce bias and facilitate the identification of different forms of hate speech.

Because the model is co-regulatory, the second step deals with notification of violations. OSPs should make it possible for law-enforcement agencies to notify them of violations, publish guidelines

directed to law enforcement, and create national contact points designed to channel priority notifications. At the same time, OSPs should also strengthen their work with civil society organizations that function as “trusted reporters” and create user interfaces for submitting complaints. These interfaces should require granular information and be located on the platform’s main user interface.

The third step deals with the organizational decision about the flagged content. After containment of the content pending a final decision, the extent of the restriction and the response time should be a function of the origin of the request. OSPs should use the common criteria to help identify hate speech, followed by a differential response to the content based on its severity. Subsequent to the decision that the content does in fact violate its policies, an OSP should notify the agency or person who filed the complaint as to its decision. Depending on the severity of the content and the company’s decision, the OSP should provide users whose content was blocked or removed with information about the decision, including whether they are entitled to appeal the decision and how to do so.

The last step aims to provide transparency and accountability. First of all, in order to maintain trust and reliance OSPs should provide users with a thorough explanation of the criteria they implement. Management should ensure that all complaints and requests are monitored and analyzed on a monthly basis. This includes the collection of data on the relevant posts, their shareability, and the decisions made. The decisions taken should be available to the relevant OSP staff in the form of detailed case studies and to the public in the form of a transparency report and open data. Additional accountability measures include counseling and support programs for content moderators and reviewers, collaboration with civil society organizations, cooperation among OSPs, reassessment of the policies by senior management, and education of users to raise their awareness about the types of content that are not permitted under the OSP’s rules and community standards.

Hate Speech on Social Media

Internet platforms and social media have a tremendous positive influence on the human ability to exchange information and ideas, to learn, to build communities and bring people together, and to promote social justice and democracy. At the same time, though, we are also beginning to see the scope of the negative phenomena that accompany these innovations—from disinformation and fake news, through infringement of privacy, mass surveillance, harmful psychological side-effects, and influencing of elections, and on to the accumulation of wealth and political power that results from control of the public discourse by a handful of persons; and, finally, hate speech and the dissemination of hatred for groups and individuals.

This policy paper addresses ways of dealing with hate speech on social media. As the dimensions of that phenomenon have become clearer, increasing thought is being given to ways of countering it.

The monitoring of hate speech on social media is inadequate. The various actors employ different methodologies in order to understand the scope of the phenomenon of hate speech on social media. Among other things, it is possible to identify attempts to quantify the posts on blogs and leading platforms such as Facebook, Instagram, YouTube, and Twitter. The World Jewish Congress, along with Vigo Social Intelligence,¹ is attempting to count the daily volume and source of antisemitic neo-Nazi posts on blogs and leading platforms. The Anti-Discrimination League (ADL) has developed a set of keywords for identifying antisemitic language on Twitter and studying how many such tweets there are, to whom they are addressed, and how other Web surfers react to them. The Pew Research Center, which focuses on the American market, employs both content analysis and surveys in order to determine whether Americans are more likely to be exposed to

1 The World Jewish Congress in collaboration with Vigo Social Intelligence, *The Rise of Anti-Semitism on Social Media: Summary of 2016*.

racist content than to publish such content. The European Union Agency for Fundamental Rights has noted the dearth of information about antisemitic utterances on social media in Europe; the information that is available is published without methodological harmonization among the EU member states. This problem may make it difficult for law-enforcement agencies and the courts to deal with the phenomenon and develop a data-driven policy for doing so.

Others are involved with hate speech in the context of specific countries, such as South Africa and Israel.² In the report of the Code of Conduct on Countering Illegal Hate Speech Online,³ and implementation of Framework Decision 2008/913/JHA in online contexts,⁴ published in 2016, 31 organizations and three public authorities reported on 2,575 items that violate the law in various countries that implement the European rules for prevention of online hate speech. A broad analysis of these data and the current situation can be found in Appendix A.

The rise in the volume of hate speech on social media has negative social implications. It is clear that the challenge posed by the need to balance the right to free speech against the prevention of hate speech directed against individuals and groups was not born with social-media platforms. However, the leveling of hierarchies and easy access to a public megaphone have engendered a significant increase in hate speech, with

² Citizen Research Centre, *Xenophobia on Social Media in SA, 2011–2017, Anatomy of an Incident: Violence in Gauteng and the "March against Immigrants"* (March 15, 2017); Berl Katznelson Foundation, *Report on Hate against Government Institutions and Democracy* (December 3, 2017) [in Hebrew].

³ IP/16/1937, European Commission – press release, *European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech*, Brussels, May 31, 2016. See *Code of Conduct on Countering Illegal Hate Speech Online*.

⁴ EU Council Framework Decision 2008/913/JHA (3) on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law.

the advent of individual players, groups, and countries that spew hate speech into the mainstream of the public domain, without the mediation mechanisms that characterized the establishment media, and with no state supervision. Second, the algorithms employed by social-media platforms and the business models of the companies that control them have created patterns of virality that allow hate speech to spread rapidly and reach extremely broad audiences; to be directed and targeted against groups and individuals (both by the platforms themselves and by those who misuse them), thereby multiplying the damage it does (because of the injury to those it targets as well as the recruitment of support for hate speech); and to be sold through content distribution services to organizations and countries that are interested in disseminating it. The data in Appendix A show an increasing trend among those motivated by intolerance, a scarcity of liberal positions, and proliferation of extremist views as a result of exposure to hate speech online. To this must be added the attempts to chalk up geopolitical profit by promoting hate as part of election campaigns in several democratic countries.

Along with the negative implications in the general social sense, online hate speech has a negative impact on individuals. Whereas it is possible to toss harassing letters into the wastepaper basket, in the digital realm nothing is ever forgotten; in fact, the harassing content can spread exponentially to various target audiences.⁵ On the psychological level, research has shown that the increase in hate speech on social media is a consistent and deliberate cause of emotional distress, because it is continual and not an isolated or one-time action.⁶ These phenomena damage individuals' work environment and good name and may even lead to physical harm, whether self-inflicted or by others.

5 DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 5 (2014).

6 *See id.*

So everyone agrees that it is essential to find solutions to the phenomenon of online hate speech. However, these solutions must take account of the fear that they would have an excessive impact on the right of free expression. Consequently, every solution must take into account the need for proportionality with regard to substance—what should be defined as hate speech, treated as such, and marked as unacceptable—but also the need for proportionality on the institutional plane—who or what is the appropriate body to decide on the rules and then to enforce them.

The present policy paper surveys the different facets of the regulation of hate speech. It does not delve into constitutional issues and seeks only to offer feasible solutions for practical implementation.

We will offer substantive definitions of hate speech as well as several forms of implementation arrangements for coping with hateful content on social-media platforms.

Defining hate speech is a complex task. There are different types of definitions both in national legislation and its implementation by courts, and in supranational legislation and international conventions and their enforcement by international panels, civil society organizations, academia, and technology companies.

On the institutional level there are attempts to enact national legislation or to apply existing national rules to the digital space, as well as international conventions and action by supranational bodies such as the European Union. There have also been attempts at self-regulation by OSPs,⁷ which have drafted organizational policies to cope with hate speech on their platforms.

⁷ There are two types of service providers: internet service providers (ISPs), which connect users to the internet, and OSPs, which users access after connecting. AT&T, Comcast Xfinity, and TimeWarner Cable are examples of well-known American ISPs. When we refer to the actual services provided by an OSP we call them "social-media platforms" or "platforms."

On the substantive level this involves defining basic rules about what is considered to be hate speech and defining sanctions for the violation of these rules as part of the platform's terms of service. Alternatively, it may involve defining rules that are an interpretation of national legislation and their application to users in that country only.

On the practical level this means the assignment of human content moderators and the development of automated algorithms to identify problematic expressions. This policy is applied, at the company's discretion, to all users of its platform throughout the world or at the state level only (a practice known as geo-blocking), based on identification of users' IP address, so as to block access by users in a specific country to content that is considered to be offensive or unlawful in that country.

The qualms associated with OSPs' self-regulatory policies are linked to the perception of regulation as a form of censorship, in this case practiced by profit-oriented companies and in a procedure that is not always transparent or democratic. State regulation and the use of geo-blocking raises the concern of regulatory islands, meaning that problematic content may be removed in one country but not in others, as well as the possibility of technological workarounds that permit access to the content even by users in a country where it is banned.

The preferred option presented in this study is one of self-regulation, but of a form that is closer and more precise than what currently exists on the internet. The self-regulation we propose includes both a content aspect and an institutional and practical aspect.

With regard to content, the definitions consist of various subsidiary definitions that are elements of what can be seen as the common definitions of hate speech in most Western countries and international conventions. We have located each of our definitions on a scale that makes it possible to choose among a range of possibilities, from the most limited to the broadest.

We propose that each platform consolidate its own policy, based on the position it deems appropriate on each scale. These choices, taken together, will constitute the platform's policy. As we see it, this will produce more precise definitions than those employed today, better reflect the general postulates of the civilized world, permit maximum transparency of policies, and make it possible for them to be applied both by human beings and by machines.

On the practical level, our recommendation seek to permit the combination of flagging of problematic content by web surfers with official notification channels for state authorities and designated organizations. This is more or less what is currently done on the large social-media platforms, but the proposed model is sufficiently flexible for it to be implemented by smaller companies as well. In addition, countries that wish to adopt a co-regulatory model will be able to draw on it. The model also includes principles of procedural transparency that we consider to be essential for its success.

In a co-regulation mechanism, OSPs and law-enforcement agencies share responsibility. The proposal draws on the OSPs' strong interest in self-regulation as an alternative to public regulation. A co-regulation mechanism for countering illegal speech, as already exists between the European Commission and OSPs,⁸ can provide the member states and OSPs with clear and accepted methods and procedures. Unlike these co-regulatory mechanisms, our model includes a clearer definition of the substantive criteria that OSPs must implement as well as detailed ways for OSPs to implement these criteria. In contrast to previous attempts, our use of scales permits OSPs to incorporate both human-based and algorithm-based mechanisms and to decide how to act when confronted by a political backlash or economic considerations.

⁸ European Commission, *supra* note 3. See Code of Conduct, *supra* note 3.

Chapter 1

The Legal Framework: International and National Interpretations of Hate Speech

In this chapter we survey the general legal framework for dealing with hate speech, as found in international and local conventions and in several Western countries. All of the documents we cite endeavor to balance the right to free expression with the public interest and with the right of individuals and groups to be protected against behavior or speech that could be interpreted as hate speech, incitement to violence, or racism. An extensive legal analysis would go far beyond the limits of this paper. Other papers in this project attempt to broaden this scope.

Our goal in this chapter is to present the fundamental principles for defining and dealing with hate speech, which will subsequently be broken down into the subsidiary definitions of our recommendations.

1.1 Global International Conventions

1.1.1. The 1948 Universal Declaration of Human Rights (UDHR)

1.1.1.1. The UDHR establishes the right to equal protection under the law. Though the UDHR has become customary international law over the years, it is not binding.

1.1.1.2. Article 7 of the UDHR states that “[a]ll are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination.”⁹

⁹ [The Universal Declaration of Human Rights \(UDHR\)](#), adopted on December 10, 1948, General Assembly resolution 217 A.

1.1.1.3. Article 19 of the UDHR states that the right of free expression includes the “freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.”

1.1.2. The International Covenant on Civil and Political Rights (ICCPR)¹⁰

1.1.2.1. According to Article 19 of the ICCPR, “[e]veryone shall have the right to hold opinions without interference” and “[e]veryone shall have the right to freedom of expression.”

1.1.2.2. This article may conflict with Article 20, which states that “[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.”

1.1.3. The Rabat Plan of Action

1.1.3.1. One attempt to balance these two articles of the ICCPR was made by the UN Office of the High Commissioner of Human Rights (OHCHR) in the Rabat Plan of Action on the prohibition of “national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.”¹¹

1.1.3.2. According to the Rabat Plan, a six-part threshold test¹² makes it possible to assess when speech is severe enough to warrant punishment under Article 20.

¹⁰ [International Covenant on Civil and Political Rights \(ICCPR\)](#), adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI) of December 16, 1966, entry into force March 23, 1976, in accordance with Article 49 (hereinafter: ICCPR).

¹¹ [Conclusions and recommendations emanating from the four-regional expert workshop organized by the OHCHR in 2011](#) and adopted by experts in Rabat, Morocco, on October 5, 2012.

¹² (1) The social and political context of the statement being made; (2) the social status or position of the speaker; (3) the specific intent to cause harm; (4) the degree to which the content is “provocative and

1.1.4. The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)¹³

1.1.4.1. The ICERD differs from the ICCPR in three ways:¹⁴

1.1.4.1.1. The ICERD is limited to hate speech that refers to race and ethnicity.

1.1.4.1.2. Article 4 of the ICERD imposes a stricter obligation on state parties.

1.1.4.1.3. The ICCPR and ICERD differ regarding intent.¹⁵

1.1.5. Other conventions and treaties that deal with more specific issues include the Convention on the Prevention and Punishment of the Crime of Genocide (1951) and the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) (1981).¹⁶

direct" and the "nature of the arguments deployed in the speech"; (5) the extent, reach, and size of the audience; (6) the likelihood that the speech will effectively incite harm (ibid.).

¹³ The Convention was adopted by the UN General Assembly in 1965 and came into force in 1969: ICCPR, *supra* note 10.

¹⁴ See IGINIO GAGLIARDONE, DANIT GAL, THIAGO ALVES, & GABRIELA MARTINEZ, *COUNTERING ONLINE HATE SPEECH* (2015), at 21–23. (hereinafter: UNESCO – Countering Online Hate Speech).

¹⁵ *Id.* at 21.

¹⁶ The Convention on the Prevention and Punishment of the Crime of Genocide is limited to public incitement of hate crimes against groups based on race, nationality, ethnicity, and religion. See UNESCO – Countering Online Hate Speech, *supra* note 14.

1.2 Regional Conventions

1.2.1. There are several regional conventions that complement the global treaties.¹⁷ For instance, both the European Convention on Human Rights and the European Union's Charter of Fundamental Rights enshrine the right to life, human dignity, equal treatment, and freedom of thought, conscience, and religion as universal human rights.¹⁸ While addressing each of these conventions and its influence on human rights online far exceeds the scope of this paper, the next paragraphs specifically address the influence of the European conventions and legislation on freedom of expression online.

1.2.2. The European Convention on Human Rights

1.2.2.1. Article 10.1 of the European Convention on Human Rights grants the right to freedom of expression, including the freedom to hold opinions and to receive and impart information and ideas without interference by public authority.

1.2.2.2. Article 10.2 states that given the duties and responsibilities derived from these freedoms, the exercise of these freedoms may be subject to formalities, conditions, restrictions or penalties that are prescribed by law and, among others, are necessary in a democratic society or for the prevention of crime or unrest.

1.2.2.3. Two rulings by the European Court of Human Rights (ECHR) apply Article 10 to online news portals. In both cases, even though the online news portals were not aware of the relevant comments,

¹⁷ See other papers in this project.

¹⁸ The protection and promotion of these rights are intimately linked with the fight against hate crimes such as antisemitism.

national courts had found them liable for comments posted on their websites.¹⁹

1.2.2.3.1. In *Delfi AC v. Estonia*, the Grand Chamber of the ECHR dealt with threats and antisemitic slurs that were published in *Delfi*, an Estonian online newspaper.²⁰ The Grand Chamber affirmed the Estonian court's decision that the platform could be liable for the comments, even though its practice was to remove such comments as soon as it found out about them. The Grand Chamber found that strict liability for users' comments does not violate the rights provided by Article 10 of the Convention, including the right to seek and impart information.

1.2.2.3.2. In *MTE v. Hungary*,²¹ on the other hand, the ECHR Grand Chamber overruled a national court decision that held the platform liable for readers' comments about the misleading business practices of two real-estate websites. The ECHR found that, in principle, an internet news portal had duties and responsibilities with regard to the comments of users – whether identified or anonymous – who engage in clearly unlawful speech which infringes the personal rights of others and amounts to hate speech and incitement to violence against them (although they are not the publishers of the comments in the traditional sense).

19 Daphne Keller, *Litigating Platform Liability in Europe: New Human Rights Case Law in the Real World*, The Center for Internet and Society (April 13, 2016).

20 *Delfi AS v. Estonia*, Application no. 64569/09.

21 Magyar Tartalomszolgáltatók Egyesülete (MTE) is a self-regulatory body; Index.hu Zrt is the owner of one of the major Hungarian internet news portals. See *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary* (Application no. 22947/13).

1.2.2.3.3. However, in *MTE v. Hungary* the ECHR found that the Hungarian courts had failed to properly balance the competing rights involved, and mainly the applicants' right to freedom of expression and the real-estate websites' right to respect for their commercial reputation. Unlike in *Delfi AS v. Estonia*, here the ECHR found that the applicants' case lacked the pivotal elements of *Delfi*: the comments might have been offensive and vulgar, but were not clearly unlawful speech in the category of hate speech and incitement to violence.²² As such, the Hungarian court's ruling violated Article 10.

1.2.2.4. Although in *Delfi* the ECHR limited its decision to the particular defendant, the result of the two cases is that platforms are required to monitor and delete comments in order to avoid liability. While compelling a platform to find and remove every unlawful user comment is an excessive and impracticable requirement that can undermine the right to impart information on the internet, it seems that the ECHR identified platforms' duties and responsibilities at least for hate speech and direct threats.

1.2.3. The Council of Europe's Cybercrime Convention and its Additional Protocol²³

22 The ECHR used the following criteria, established in case law for the assessment of proportionality of the interference in situations not involving hate speech: the context and content of the comments, the liability of the authors of the comments, the steps taken by the applicants and conduct of the injured party, and consequences of the comments.

23 The Council of Europe, *Convention on Cybercrime*, opened for signature on November 23, 2001, entered into force on July 1, 2004 (ETS no. 185). Council of Europe, *Additional Protocol to the Convention of Cybercrime*, concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems, opened for

1.2.3.1. The Cybercrime Convention facilitates cooperation between countries in combating computer-based crimes; the Additional Protocol covers online hate speech.

1.2.3.2. The Additional Protocol calls for the criminalization of the dissemination of racist and xenophobic materials, threats, and insults via computer systems.²⁴

1.2.3.3. The Additional Protocol also covers the denial and justification of genocide and crimes against humanity and provides for the extradition of hate-speech offenders.

1.2.4. Several European directives address discrimination on ethnic or racial grounds.²⁵ Most notable here is Council Framework Decision 2008/913/JHA (November 28, 2008) on the use of criminal law to combat certain forms and expressions of racism and xenophobia.

1.2.4.1. Decision 2008/913/JHA seeks to define a standard EU-wide criminal-law approach to countering severe manifestations

signature on January 28, 2003, entered into force on March 1, 2006 (ETS no. 189).

24 *Id.*

25 The Racial Equality Directive (2004/43/EC) prohibits discrimination on the grounds of racial or ethnic origin in employment; the Employment Equality Directive (2000/78/EC) prohibits discrimination in employment on the grounds of religion or belief. The Victims' Rights Directive (2012/29/EU) establishes minimum standards for the rights, support, and protection of victims of crime. It refers explicitly to victims of hate crimes, their protection, and the specific needs related to their recognition, respectful treatment, support, and access to justice.

of racism and xenophobia.²⁶ It contains no binding provisions, however.²⁷

1.2.4.2. In the attempt to ensure that certain behaviors constitute an offense in all EU member states, Decision 2008/913/JHA defines hate speech as one of three actions:²⁸

1.2.4.2.1. Public incitement to violence or hatred directed against a group of persons or a member of such a group, defined by reference to race, color, religion, descent, or national or ethnic origin;

1.2.4.2.2. The same, when done through the “public dissemination or distribution of tracts, pictures or other material”;

1.2.4.2.3. “Publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity, and war crimes [as defined in EU law], when the conduct is carried out in a manner likely to incite violence or hatred against such a group or a member of such a group.”²⁹

26 Council Framework Decision 2008/913/JHA of November 28, 2008, on combating certain forms and expressions of racism and xenophobia by means of criminal law sets out to define a common EU-wide criminal-law approach to countering severe manifestations of racism.

27 Its goal is to indicate how relevant EU and member-state laws should be interpreted. See Andrew F. Sellars, *Defining Hate Speech* (December 8, 2016), Berkman Klein Center Research Publication No. 2016–20.

28 It also requires member states to provide effective, proportionate, and dissuasive criminal penalties (including the possibility of imprisonment) for natural and legal persons who have committed or who are liable for offenses motivated by racism or xenophobia, including antisemitism.

29 See Article 1 of Decision 2008/913/JHA. See also the summary in Sellars, *supra* note 27, at 20.

1.3 National Implementation

1.3.1. These supranational attempts to harmonize the definition of hate crimes and to balance it with other human freedoms may be applied differently at the national level.³⁰ In addition, the conditions for determining jurisdiction may vary from country to country.³¹

1.3.2. Where national legislation places the content within the country's jurisdiction, this may result in the implementation of several policy instruments, as detailed in Chapter 3.

1.3.3. The United States and Europe offer distinct perspectives in many ways:

1.3.3.1. Whereas the U.S. does not make hate speech per se illegal under any definition, the German and French systems are stricter, due to their cultural heritage and for historical reasons.³²

1.3.3.2. In the U.S., the legal system uses defamation laws to protect people's reputations. The courts can create, balance, and limit First

30 Usually, each country's criminal code determines when a specific statement is considered to have been made on its territory, so that the act or statement falls under its jurisdiction.

31 These may include: (1) the place where the instigator uploaded the content; (2) the instigator's citizenship status; (3) the victim's citizenship status; (4) where the content is accessible; (5) the place from which the content was made available; (6) whether the content targets the country's citizens. States can also claim jurisdiction over online hate speech based on (7) the location of the server and (8) whether the content is accessible to its citizens. See Talia Naamat & Elena Pesina (2016), [Legislation Survey: Regulating Online Hate Speech in Europe](#), p. 3. Kantor Center for the Study of Contemporary European Jewry (hereinafter: Kantor Center – Legislation Survey).

32 Sellars, *supra* note 27, at 5; James Q. Whitman, *Enforcing Civility and Respect: Three Societies*, 109 YALE L. REV. 1279 (2000).

Amendment doctrines.³³ Subjectivity and elusive definitions are a consequence of the American approach.³⁴

1.3.3.3. The French Penal Code punishes hate speech with five years' imprisonment and a fine of 300,000 Euros.³⁵ According to the Press Freedom Law, hate speech is punishable by five years' imprisonment, a fine of 45,000 Euros, or both only if the incitement did not lead to effective action.³⁶

1.3.3.4. In Germany, the Criminal Code prohibits incitement to hatred through written materials, including media storage and audiovisual media. Incitement to hatred is punishable by three months to five years' imprisonment; the dissemination or public display of hate speech can lead to imprisonment of up to three years or a fine.³⁷ In addition, ISPs are required to provide customer details to the public prosecutor upon request, and the German Telecommunications Law allows the storage of IP addresses if the offense was committed via telecommunication services.³⁸ However, unlike France, Germany has no online mechanism for the submission of reports about hate speech content.

33 See also James Banks, *Regulating Hate Speech Online*, INTERNATIONAL REVIEW OF LAW, COMPUTERS & TECHNOLOGY 24:3 (2010), at 233. See further explanation in other papers in this project.

34 Sellars, *supra* note 27, at 5–8.

35 Article 226–19 of the Penal Code, Article 24 and 24bis of the Law on Press Freedom in Kantor Center – Legislation Survey, *supra* note 31, at 40.

36 Article 24 and 24bis of the Law on Press Freedom; see Kantor Center – Legislation Survey, *supra* note 31, at 40.

37 Sections 11, 130, 130a, 131 of the Federal Criminal Code; see Kantor Center – Legislation Survey, *supra* note 31, at 47.

38 See Kantor Center – Legislation Survey, *supra* note 31, at 50.

1.3.3.5. The Canadian criminal code punishes anyone who “willfully promotes hatred against any identifiable group” but excludes various types of statements.³⁹ Canada also prohibits public statements that incite hatred against any identifiable group if that statement is likely to lead to a breach of the peace⁴⁰ or is an advocacy for, or promotion of genocide.⁴¹

1.3.3.6. In the United Kingdom, the Public Order Act of 1986 prohibits the dissemination or display of speech that is “threatening, abusive or insulting,” if the speaker intends to stir up racial hatred or if “having regard to all the circumstances racial hatred is likely to be stirred up thereby.”⁴² This rule applies to both deliberate speech and consequential harm, as well as to negligence.⁴³

39 These exclusions include statements that are proven by the defendant to be true, statements that are offered “in good faith” or when expressing “an opinion on a religious subject,” statements that are “relevant to the public interest, the discussion of which was for the public benefit,” or if “in good faith,” the person was pointing out other hate speech “for the purpose of removal.” Canada Criminal Code §319(3). See also Sellars, *supra* note 27, at 19.

40 *Id.* §319(1).

41 Targeted groups can include groups identified by color, race, religion, national or ethnic origin, age, sex, sexual orientation, or mental or physical disability. *Id.* §318.

42 United Kingdom Public Order Act 1986 §18(1).

43 Sellars, *supra* note 27, at 19.

Chapter 2

The Different Categories of Players and the Responsibility of Internet Intermediaries

Any discussion about the regulation of hate speech on social-media platforms must consider the several players involved. A first typology relates to those implicated by the speech itself. Here we find the content originator (or aggressor), the objects of the publication (the individuals or groups whom the hate speech attacks), and other actors (those whom the originator wishes to persuade, those who share or “like” the content). We will deal with these mainly in the context of the substantive definitions of hate speech and when we address the question of the platforms’ obligation to block virality, that is, to keep content from reaching additional audiences, having additional shares, and so on.

A second typology relates to the actors involved in regulation. Here we can list state actors (governments, law-enforcement agencies), supranational actors (international organizations such as the United Nations and the European Union), civil society and consumer organizations, and finally companies that develop technological solutions for applying regulations.

A third typology relates to the platforms on which the hate speech is posted. These platforms can be:

- (1) Social-media platforms that are open to the general public; that is, they require registration and identification, but after users enter them they make the content available to the public at large: Facebook, Twitter, Instagram, Gab, and so on.
- (2) Social-media platforms for defined groups—WhatsApp groups, Telegram groups, closed Facebook groups. These groups require registration and their content is open only to members of the group and not the public at large.
- (3) Hosting services for content sites that are intended for the general public, such as blogging platforms that provide only technical support—

GoDaddy, WordPress, and Reddit, and dedicated blogging platforms such as Blogger, Tumblr, and Medium.

(4) Closed hosting services that allow individuals and companies to store data online, such as the cloud services run by Microsoft and Amazon.

OSPs are also known as “content intermediaries.” An intermediary is the means by which information is conveyed from one side to another. According to the OECD definition, internet intermediaries bring together or facilitate transactions between third parties on the internet. They give access to, host, transmit and index content, products and services originated by third parties on the internet or provide internet-based services to third parties. This definition leaves out independently created content on sites that are pre-edited, such as Wikipedia and traditional news sites, as well as content sites and blogs located on private domains (that is, not on hosted sites), subscription television services, and the like. In any case, an intermediary does not fall into the category of “the media,” because the primary condition for defining a content site as a journalistic media channel is the exercise of editorial discretion and adherence to professional and ethical standards.

On the surface, the fact that social-media platforms have terms of service that govern content, which users are required to accept, means that they too have editorial discretion about content. However, these terms of service are associated with contract and commercial law rather than the fields of media regulation, communications law, and freedom of expression and freedom of the press.

The standard definition of intermediaries thus refers to companies that host, provide access to, index, promote, or permit the transfer or sharing of content created by others. Intermediaries can be categorized by the technical function or role they play. Of course, the several categories of intermediaries have different business models and different geographical locations, employ different technologies, and are subject to different legal

regimes. By the same token, states' ability to limit expression varies among the different types of intermediaries. For example, a state can block ISPs and thereby prevent its citizens from accessing the internet, or it can block access to a particular intermediary that provides a specific service. This study deals only with the first three categories of platforms. As we see it, closed hosting services do not have the same negative social impact as sites with content that is intended for the public or groups, whether defined or not. Finally, sites that practice content-editing in any case base their decisions upon the residence of the content creators, who can be located easily and subjected to legal provisions according to a geographic key.

Many OSPs are multinational entities that provide social-network platforms for transnational markets, and their operations transcend national borders. This characteristic does not eliminate their obligation to implement each country's relevant legislation regarding users in a particular jurisdiction. Specifically, as explained, the definition of hate crimes varies widely from state to state. However, there is also a significant difference among countries when it comes to online intermediaries' exemption from liability for content published on their platforms by their users. On the one hand, this immunity facilitates innovation on social-media platforms and their development as an important public arena. On the other hand, the rules on platform liability, and more importantly the exceptions to those rules, affect the intensity of the monitoring that OSPs must devote to preventing the use of their platforms for illegal activities and speech.

An examination of the legal situation of internet intermediaries in different countries reveals that there are three main models. The first is that of strict liability, which holds the intermediary responsible for all content on its platform and liable for third-party content unless it has established a mechanism to screen, monitor, and delete content. The second model is that of conditional liability, which relieves the intermediary of liability for third-party content if certain conditions are met; for example, if the intermediary deletes content when it receives notice to do so ("notice and takedown"), if it informs the content creator that it has received

a warning about the legality of the content (“notice and notice”), or if it disconnects repeat offenders. The third model is that of broad immunity for intermediaries for all third-party content.

According to Tarleton Gillespie, these liability rules for online intermediaries pose three challenges.⁴⁴ First, the platform-liability laws were originally designed in the era of ISPs, homepages, and online community discussion forums, and not for the digital economy and the platform capitalism era.⁴⁵ Second, much like the laws that criminalize hate crimes, the platform liability rules are country-specific; but many and especially the largest service providers are multinational corporations that operate simultaneously in several jurisdictions. This second challenge, in turn, corresponds to the third challenge—the difference between jurisdictions as to the extent of the liability a platform faces, and on what grounds. Above we looked at the differences in the laws on hate speech and racial discrimination, but there are also different interpretations of copyright infringement, reactions to cybercrime and terrorist content, and definitions of legitimate speech or socially acceptable content. The contrasting American and European laws exemplify the different immunity regimes that national legislation grants platforms. Two other forms of intermediary liability, which are not discussed here, are countries with “strict liability” regimes, which require providers to proactively prevent or censor the circulation of illicit or unlawful content (China is the leading example), and countries with no intermediary liability laws.⁴⁶

⁴⁴ Tarleton Gillespie, *Regulation of and by Platforms*, THE SAGE HANDBOOK OF SOCIAL MEDIA (J. Burgess, A. Marwick, & T. Poell, eds., 2018).

⁴⁵ Platform capitalism means an economy based on OSPs that provide others (consumers and producers) with the hardware and software foundations to operate on.

⁴⁶ REBECCA MACKINNON, ELONNAT HICKOK, ALLON BAR, & HAE-IN LIM, [FOSTERING FREEDOM ONLINE: THE ROLES, CHALLENGES AND OBSTACLES OF INTERNET INTERMEDIARIES](#) (2014), at 40. (hereinafter: UNESCO – Fostering Freedom Online); Gillespie, *supra* note 44, at 6.

In the United States,⁴⁷ Section 230 of the Communication Decency Act (CDA) states that an “interactive computer service provider” cannot be held liable for content published by users. The reasoning behind this immunity is that the provider merely provides access to the internet and other services.⁴⁸ Section 230 exempts a platform-provider that claims to be “an interactive computer service” from being treated as a publisher of information or content. However, this exception has a secondary clause, known as the “Good Samaritan” rule. The first rule does not require providers to police their users. But if the provider decides to do so anyway, the second rule comes into effect: the provider is still not deemed to be the publisher of the content and remains immune to liability.⁴⁹ The goal of this second rule is to avoid discouraging providers from policing content, as would occur were their liability reinstated as the result of a decision to intervene and police content on their platform. In fact, according to Gillespie, nearly all platform operators impose their own rules and monitor offensive content and behavior on their platforms. Because platforms are not government actors, they are not required to protect their users’ speech under the First Amendment,⁵⁰ though legal scholars tend to demand this protection from the OSPs.⁵¹

47 The constitutional implications of Section 230 of the CDA far exceed the scope of this paper. The following paper in our project deals more broadly with these issues. See Karen Eltis and Ilia Maria Siatitsa, *Realigning the Law to Better Uphold the State's Duty to Protect Human Rights: Towards an Interoperable Model for Addressing Racism and Strengthening Democratic Legitimacy*.

48 47 U.S.C. §230(c)(1).

49 47 U.S.C. §230(c)(2).

50 Sellars, *supra* note 27, at 21.

51 Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

There have been attempts to chip away at Section 230 of the CDA, based on the claim that platforms solicit or structure unlawful behavior through their user interface and thus help to foster illegal content.⁵² For instance, in *Fair Housing Council of San Fernando Valley v. Roommates.com*,⁵³ the Ninth Circuit Court of Appeals found that the listing service Roommates.com was not entitled to the CDA immunity, because its drop-down menus were structured to facilitate users' entry of discriminatory preferences about roommates. That is, the platform made discriminatory questions part of "doing business" on the website.⁵⁴ The *Roommates.com* decision produced extensive legal scholarship about how it affects or limits the Section 230 immunity and made design decisions a factor in the regulation of users' conduct.

Nevertheless, despite the attempts to reduce the platforms' broad immunity, their business models have enabled them to sidestep the traditional rules aimed at preventing discrimination. In addition, platforms' terms of use include a disclaimer of liability when users assert damage caused by other users.⁵⁵ As such, Section 230 immunity is the

52 Several recent cases directly address the liability of Facebook, Google, and Twitter for failing to prevent foreign terrorist organizations from using their social-media platforms. The courts, for the most part, upheld Section 230 protection. However, the 9th Circuit in *Fields v. Twitter* found that plaintiffs can show that the social-media sites had a "direct relationship" to the terrorist attacks (the higher proximate causation standard). See *Fields v. Twitter, Inc.*, 2018 WL 626800 (9th Cir. Jan. 31, 2018). In these cases, the plaintiff attempted to claim, for instance, that YouTube shared revenues with the terrorists. See, e.g., *Gonzalez v. Google, Inc.*, 2018 WL 3872781 (N.D. Cal. Aug. 15, 2018). See also Eric Goldman, *The Ten Most Important Section 230 Rulings*, 20 TUL. J. TECH. & INTELL. PROP. 1 (2017); for further cases, see [Eric Goldman's blog](#).

53 *Fair Housing Council of San Fernando Valley v. Roommates.com*, 521 F.3d 1157, 1168 (9th Cir. 2008).

54 *Id.* at 1181.

55 See Orly Lobel, *The Law of the Platform*, 101 MINN. L. REV. 87 (2016). See also Karen Levi & Solon Barocas, *Designing against Discrimination in Online Markets*, 32 BERKELEY TECH. L. J. 1183, 1187 (2017).

most lenient of all intermediary liability regimes and is termed “broad immunity.”⁵⁶

In Europe, by contrast, Directive 2000/31/EC harmonizes the member states’ legislation on e-commerce and provides that internet intermediaries will not be held liable if their actions satisfy certain conditions.⁵⁷ Such “conditional liability”⁵⁸ exists in the United States as well under the Digital Millennium Copyright Act.⁵⁹ According to Article 12 of Directive 2000/31/EC, internet intermediaries are not required to actively monitor information and content stored on their servers or platforms. The result is that internet intermediaries have no incentive to install self-monitoring mechanisms. However, when an internet intermediary is notified of illegal content and thus receives “actual knowledge” of the problematic content, it must block access to or remove the content. The timeframe for content removal varies from country to country—“expeditiously,” “within a reasonable time,” “immediately,” “24 hours.”⁶⁰ Failure to remove the content may lead to administrative or civil liability.

56 UNESCO – Fostering Freedom Online, *supra* note 46, at 42; Gillespie, *supra* note 44, at 6.

57 Directive 2000/31/EC of the European Parliament and the Council, June 8, 2000, on certain legal aspects of information society services, in particular electronic commerce, in the internal market.

58 UNESCO – Fostering Freedom Online, *supra* note 46, at 40; Gillespie, *supra* note 44, at 6–7.

59 Under the Digital Millennium Copyright Act’s conditional liability (known also as notice-and-takedown), service providers are not liable for what their users have uploaded or distributed as long as they have no “actual” knowledge of the content and did not produce or copy the illegal or illicit materials. Service providers need also respond to requests by copyright owners who identify their work as circulating through the platform. Material contribution to the circulation of pirated content, financial benefits from it, or promotion of the service as designated for privacy can take away the exemption from liability.

60 Kantor Center – Legislation Survey, *supra* note 31, at 4.

The European Court of Justice has addressed the issue of the liability of online service providers. Most cases relate to matters of data-protection violations and infringement of intellectual property rights.⁶¹ Among them, one case relates to social-media platforms. In *SABAM v. Netlog*, the European Court of Justice found that a Belgian court could not require Netlog to install a filtering system that would conduct active monitoring of all user data and prevent future infringements of intellectual property.⁶²

61 Well-known cases on data-protection violations and intellectual-property infringement that limited the scope of the exemption from liability include: *Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos and Mario Costeja Gonzalez*, C-131/12 (finding that people have the right to be forgotten on search engines); *GS Media BV v. Sanoma Media Netherlands BV and Others*, Case C-160/15 (finding rebuttable presumption of knowledge in cases of links made for profit); *L'Oréal SA and Others v. eBay International AG and Others*, C-324/09 (providing clarifications for OSPs' liability for trademark infringement committed by their users in their internet marketplace); *Nils Svensson et al. v. Retriever Sverige AB*, C-466/12 (links to authorized works freely available online do not infringe the owner's copyrights); and *ITV Broadcasting Ltd. and others v. TVCatchup Ltd.*, C-607/2011 (sites that link to streams are responsible for communicating copyrighted works to the public). Another case worth mentioning is *Schrems v. Data Protection Commissioner*, C-362/14, where an Austrian Facebook user initiated the invalidation of Commission Decision 2000/520/EC that created the transatlantic U.S.-EU Safe Harbor agreement.

62 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v. Netlog NV*, C-360/10. Similarly, in *Scarlet Extended SA v. SABAM* the court found that a collective rights-management organization could not require ISPs to install a filtering system to prevent the illegal downloading of files, as it would seriously endanger "the freedom to conduct business enjoyed by operators such as ISPs" and would possibly infringe "the fundamental rights of that ISP's customers, namely, their right to protection of their personal data and their freedom to receive or impart information." See European Court of Justice, *Scarlet Extended SA v. Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)*, C-70/10, November 24, 2011.

Content moderation for social-media platforms is, however, regulated on the member state level.

The most recent national regulation is that in Germany, where teleservice providers are not required to monitor third-party content or disconnect customers who infringe third-party rights. If, however, an ISP becomes aware of illegal content it is expected to block access to it.⁶³ The Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act), passed in July 2017, sets specific requirements and procedures to be implemented by the providers of “telemedia” services.⁶⁴ These requirements apply to multinational service providers that have more than two million registered users in the Federal Republic of Germany if their platforms are designed to enable users to share any content with other users or to make such content available.⁶⁵ According to the new law, “telemedia” service providers must follow transparency requirements and develop procedures to handle complaints about unlawful content. Content must be removed within 24 hours or one week, depending on whether or not it is manifestly unlawful. Failure to comply with the Act may be deemed a regulatory offense, incurring a fine levied by the Federal Office of Justice of up to five million euros, depending on the violation.⁶⁶

63 See [§§3 and 5 of the Teleservices Act in Kantor Center – Legislation Survey](#), *supra* note 31, at 48. Recently, Facebook, Twitter, and Google agreed with the German government to remove hate speech within 24 hours after notification.

64 “Telemedia service providers” is the translation of a German legal term originating with the German Telemedia Act.

65 Platforms with fewer than two million German registered users and platforms that offer journalistic or editorial content are exempt from the legislation. See Section 1 of the Network Enforcement Act.

66 Section 4 of the Network Enforcement Act.

In France, ISPs are required to take part in the fight against hate speech. However, ISPs and hosting services are not obliged to monitor the information they transmit or store.⁶⁷ In its 2012 decision, the Court of Cassation held that “obliging internet stakeholders to prevent any reposting of unlawful content which they have removed following due notification by users would be tantamount to subjecting them to a general duty to monitor the images they stock and to look for unlawful reproductions. This could not be accepted.”⁶⁸ In practice, there are two procedures for taking down content: administrative blocking and court orders. The authorities may order the blocking or filtering of certain sites or removal of content. To do so, they must contact the hosting service or the editor and inform the ISP of the blocking measures they ordered. Courts can require the hosting service or access provider to prevent the violation resulting from the content. If the hosting service does not comply or the administrative authority does not have the offender’s contact details, the ISP can be requested to block access.⁶⁹ A service provider has 24 hours to act;⁷⁰ failure to comply with the request is punishable by a fine of 375,000 euros and either a permanent or temporary ban of up to five years on directly or indirectly conducting professional or corporate activities.⁷¹

67 Article 6-I-7 of the Law for Confidence in the Digital Economy.

68 [Comparative Study on Blocking, Filtering and Take-Down of Illegal Internet Content](#), 2015. French Court of Cassation, Civil Division, July 12, 2012, Nos. 11-15.165, 11-13.669 and 11- 13.666. See also Kantor Center – Legislation Survey, *supra* note 31, at 41.

69 Articles 6-I-7 and 6-I-8 of the Law for Confidence in the Digital Economy. See also Kantor Center – Legislation Survey, *supra* note 31, at 41.

70 Article 6-I-1 of the Law for Confidence in the Digital Economy.

71 Whereas no civil liability is possible if there is no actual knowledge of the unlawful nature of the activity, the law determines a presumption of knowledge after the service provider receives notice.

Similarly, in Austria, intermediaries have no obligation to monitor content. After a court order is received, ISPs must provide facilities for intercepting hate speech.⁷² In addition, the Federal Agency for State Protection and Counter Terrorism may contact a service provider and ask it to inform local and international partners or providers about the violation, so that they can take action. Unlike Germany and France, Austria does not set a timeframe for content removal, but service providers are expected to act expeditiously to remove the content or block access to it.⁷³ The Austrian law also defines when an offense is considered to have been committed in Austria.⁷⁴

Unlike the United States and Europe, where there are federal and supranational laws (respectively) that define the responsibilities of content intermediaries, Israel has no analogous legislation that specifies a uniform rule for intermediaries' liability for the publication of content created by third parties. As a result of this legal lacuna, intermediaries' responsibility needs to be determined separately for each field and each case, subject to the courts' interpretation of tort law. For example, the main form of liability in Israeli law is the civil tort of negligence, defined in Sections 35 and 36 of the Torts Ordinance.⁷⁵ This text, and its interpretation by the Supreme Court, define the framework of the tort of negligence, and especially the conceptual duty of care and the concrete duty of care. In the context of liability for a third-party publication, Sections 11 and 12 of the Defamation (Prohibition) Law define the liability of advertisers, printers,

See article 4.I.2 of the Law for Confidence in the Digital Economy, and Kantor Center – Legislation Survey, *supra* note 31, at 42–43.

⁷² Austrian Telecommunications Act of 2003. *See also* Kantor Center – Legislation Survey, *supra* note 31, at 7.

⁷³ Article 16 of the Federal Act Governing Certain Legal Aspects of Electronic Commercial and Legal Transactions; *see also* Kantor Center – Legislation Survey, *supra* note 31, at 7.

⁷⁴ *See* Kantor Center – Legislation Survey, *supra* note 31, at 8–9.

⁷⁵ Torts Ordinance (new version).

and distributors for the traditional media.⁷⁶ Because of the need to update the legislation to suit the Internet Age, the courts have had to interpret these clauses to cover the liability of intermediaries, site administrators, and companies that provide platform services. Like the Defamation Law, the Protection of Privacy Law also addresses the categories of newspaper advertisements, printing, and distribution; it too was written before the Internet Age.⁷⁷ Whereas the Privacy Law stipulates that a periodical's editor, printer, and distributor may bear criminal and civil liability, it states that they will be exempt if they did not know or were not required to know that the publication constitutes an infringement of privacy.

Given the reliance on judicial interpretation, civil society's opposition to warrants issued by the police without judicial oversight, and the need to balance limitations on access to content and websites against the freedom of expression, the courts became a key element in the Israeli content-moderation process. In July 2017, for example, the Knesset passed a law that empowers district court judges to issue orders to shut down or remove or ban access to websites used to commit offenses.⁷⁸ If the conditions stipulated in the law are met, a judge can bar access to all or parts of a website or order its removal. If the website is stored outside Israeli jurisdiction, the court can order a search-engine service to prevent access to the website in question. Several Knesset committees are currently debating additional bills on the subject. At the same time, the Cybercrime Unit of the Justice Ministry employs an alternative method of enforcement and sends requests to remove content that violates Israeli laws, mostly to Facebook.⁷⁹

76 Defamation (Prohibition) Law, 5725–1965.

77 Defense of Privacy Law, 5741–1981, §§30–31.

78 Authority to Prevent Offenses by Means of a Website Law, 5717–2017.

79 This procedure was approved by the State Attorney and the Attorney General. See [Letter by the Justice Ministry, Freedom of Information Unit regarding FOIA request number 130/18](#) [in Hebrew].

In response to the different national liability laws, takedown requests, and warrants, multinational OSPs implement a policy of geo-blocking. Geo-blocking is a mechanism that originated in e-commerce, in which OSPs and online sellers deliberately restrict access to websites and content based on users' country of residence. Geo-blocking, like other practices such as geo-targeting, is based on geo-location tools that enable websites to identify an online visitor's location.⁸⁰ Geo-location has many benefits and drawbacks. It is deprecated, as by the European Union, when it is used to erect barriers in otherwise borderless environments,⁸¹ such as by online content creators and online platforms that differentiate between member states. In e-commerce, geo-location can prevent consumers from buying products that might lead retailers to run afoul of the consumer protection laws of another country; in advertising, geo-location enables retailers to localize their message. Geo-location can be used to help OSPs comply with national legislation regarding content without forcing them to delete the content or limit access to their audience worldwide.

At the same time, policymakers and OSPs alike are aware that users can circumvent geo-location measures imposed by content creators and service providers. Virtual Private Networks (VPNs) enable users to extend their network across the internet to reach servers located in other countries where the desired content is accessible, and thus to bypass territorial restrictions. Another way for users to access data is by means of web services, such as illegal streaming services, that do not employ geo-location, or through the dark web.

80 Néstor Duch-Brown & Bertin Martens, *The Economic Impact of Removing Geo-blocking Restrictions in the EU Digital Single Market*, Institute for Prospective Technological Studies Digital Economy Working Paper 2016/02, The Joint Research Center Technical Reports, The European Commission (2016).

81 *Id.*

Chapter 3

A Typology of Legal and Regulatory Instruments for Moderating Hate Speech on Social Media

Within the context of content moderation, law-enforcement agencies, ISPs, OSPs, civil society, and in some cases even users can wield different types of policy instruments. When they do so they can change the behaviors of users and the platform.

In this chapter we describe three types of content-moderation instruments: legal instruments, self-regulatory instruments, and information instruments. In each classification we identify several subgroups, in order to show the variety of options in each. After doing so we will be able to select our proposed model.

3.1 Legal and Regulatory Instruments

3.1.1. Legal instruments take the form of statutes, regulations, and court orders that require ISPs and OSPs to take certain steps or that enable law-enforcement agencies to ask providers to do so. For the most part, law-enforcement agencies implement non-contractual legal and regulatory instruments to maintain public order or to protect private interests.⁸²

3.1.2. In the next few paragraphs, we identify two groups of legal policy instruments: legislation, and court orders and warrants. We begin with

⁸² Contracts and terms of use, in this regard, are considered in this document as self-regulation and will be discussed later in the analysis.

statutes that define certain behaviors as criminal or as carrying civil liability. Then we address court orders and warrants and the actions they can instruct service providers to take.

3.1.2.1. Legislation

3.1.2.1.1. States can enact legislation that criminalizes specific behaviors, including hate speech. The statutes may further classify the offenses according to their severity: civil infractions, misdemeanors, or felonies.

3.1.2.1.2. In many cases, the legislation implements requirements set by global or supranational conventions. For instance, states that are signatories to the ICCPR are required by Article 20 to outlaw any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence. This requirement does not necessarily mandate the criminalization of all hate speech.

3.1.2.1.3. Legislation can also define civil liability, contractual or tort, for an action or inaction. For instance, in Europe, Directive 2000/31/EC sets conditions under which ISPs may enjoy immunity from liability.

3.1.2.1.4. An action that is exempt from civil liability may not receive the same treatment under the criminal code. In the United States, under Section 230 of the CDA, hate speech may not be removed unless it is also obscene, the request to take down the content is submitted by its copyright holder and is based on the copyright laws,⁸³ or the act of publication or the content itself violates federal criminal law.

83 See Digital Millennium Copyright Act, 17 U.S.C. §512.

3.1.2.2. Warrants, subpoenas, and court orders

3.1.2.2.1. Law-enforcement agencies and litigants, as part of a criminal proceeding and civil proceeding respectively, can request a court order that limits the actions of ISPs and OSPs.

3.1.2.2.2. Court orders can issue directly from a case in progress, such as a criminal investigation of hate speech; or indirectly, when law-enforcement agencies are investigating a hate crime and ask the court to limit the action of service providers or news agencies.

3.1.2.2.3. The requests can fall into several categories:

3.1.2.2.3.1. Requests to remove content

3.1.2.2.3.2. Requests to block access to websites or applications

3.1.2.2.3.3. Requests to filter content, and the “lighter” form of installing software to protect users from injurious content

3.1.2.2.3.4. Requests to disconnect users

3.1.2.2.3.5. Requests for details about a user

We now address each of these types in greater detail:

3.1.2.2.4. **Requests to remove content:** Law-enforcement agencies can request or require the deletion of questionable or illegal content. In some cases, this will be the result of a court order or of a warrant issued by a (senior) police officer. In practice, content can be removed in one of three ways:

3.1.2.2.4.1. Law-enforcement agencies can require service providers to prevent the publication of specific content, a method also known as preemption. Here the first step is usually prior identification of the content as problematic and a subsequent human

decision to remove it.⁸⁴ Another possibility is that after content has been classified as problematic or illegal, a computer can implement the decision (as discussed below for algorithm-based instruments).

3.1.2.2.4.2. Identification of the content and its subsequent removal may occur after an instigator has uploaded the content to the hosting service. After the content is flagged or identified as problematic or illegal, law-enforcement agencies can ask the ISPs or OSPs to remove it.

3.1.2.2.4.3. After the identification of content as questionable or illegal, it can be monitored across one or several platforms; this method usually involves hashing in order to save decision-making resources.⁸⁵

3.1.2.2.5. **Blocking access to websites and applications:**

Law-enforcement agencies can request or require service providers to block access to the websites or applications on which the instigator published the content. Access can be blocked in five ways:

3.1.2.2.5.1. Court orders and warrants can require ISPs to block IP addresses. Because every website must be hosted on a server, and the server has a unique

84 JONATHAN ZITTRAIN, *THE FUTURE OF THE INTERNET — AND HOW TO STOP IT* (2008).

85 *Hashing* means applying a mathematical function to a file that includes illegal content. This function creates a one-to-one identifier of the content. If a user tries to upload the content to the Internet again, the content can be monitored and blocked using the digital signature when the new file is compared by the digital system to the previous hash.

and permanent IP address, it is possible to block access to a specific IP address. Law enforcement and service providers can similarly block apps through a smartphone's or operating system's app store. A user trying to access a blocked website or app cannot connect or find the requested content.

3.1.2.2.5.2. Law-enforcement agencies can ask that websites be deregistered from national domain name system (DNS) servers.

3.1.2.2.5.3. Court orders and warrants can require ISPs to block a specific DNS server. In this case, whenever a user tries to access an unauthorized domain name, the requested DNS server will be blocked and the domain name will not be translated into an IP address, making the website unreachable. In other words, unlike IP blocking, this method blocks the web address rather than the IP.

3.1.2.2.5.4. Court orders and warrants can require the filtering of websites via an HTTP proxy. Users must transit through a proxy server that filters content before they can access it.

3.1.2.2.5.5. Court orders can also geo-block. This means that the owners of a website or service providers block access to content that is considered illegal in one or more countries, but the content is still available to users in other countries. Depending on the interests involved, either law enforcement or private actors can initiate geo-blocking. For example, law-enforcement agencies and courts usually request geo-blocking of hate speech, while private actors typically ask for geo-blocking of content protected by copyright or defend licensing

arrangements between production companies and broadcasting networks.

3.1.2.2.6. Requests that content be filtered. Such requests can be made in one of two ways:

3.1.2.2.6.1. Filtering software: ISPs can be required to install filtering software to identify prohibited content before it reaches the users/audience. Similarly, content providers can be asked or required to block access to pages that present such content.

3.1.2.2.6.2. Removal of search results: Search engines can be instructed to remove search results, change their ranking, or alter their location within the search results. Problematic or unwanted search results can be pushed down in the listing of results presented. Other service providers that store content on their services can be asked to remove the content directly.

3.1.2.2.7. Installing software to protect users:

3.1.2.2.7.1. A “lighter” form of filtering requires the installation of software to protect users (in many cases children) from harmful content. For example, users can install content filtering software or firewalls, using parental control software.

3.1.2.2.7.2. These software and services can be part of the computer’s operating system, be provided by the ISP, or be a separate software package that users acquire on request.

3.1.2.2.7.3. It is also possible to change the opt-in/out defaults of the requirement to install these filtering services. Some legislators and law enforcement may require ISPs to install filtering software, with an opt-out option for users who do not wish to have it.

3.1.2.2.8. Disconnecting users from the service or application:

3.1.2.2.8.1. This sanction means the removal of a personal, professional, or business profile from a social-media platform. For example, several countries have implemented three-strikes laws for copyright infringement.

3.1.2.2.8.2. Disconnection can involve a single platform or several ISPs. It may target a specific username, personal identifiers, or IP addresses, for a predefined period or as a permanent measure. For instance, the three-strikes policy for online copyright infringement means that if a user has been caught infringing copyright laws three times, ISPs must disconnect the user from the internet in the user's country.

3.1.2.3. Procedural and transparency measures

3.1.2.3.1. Legislation can require ISPs and OSPs to implement procedural and transparency measures. These requirements are intended to deal with the challenges presented by information and telecommunication technologies that enable individual communications and the dissemination of specific content.

3.1.2.3.2. Legislation on the implementation of procedural and transparency measures can impose reporting obligations on service providers, along with specific duties and responsibilities to handle content-removal complaints and the publication of legal notices in a defined format.

3.1.2.3.3. Legislation can also define the relevant law-enforcement agency charged with enforcing the procedural and transparency measures and empower the agency to levy administrative fines or initiate criminal proceedings.

3.1.2.3.4. A recent example of such legislation is the German Act to Improve Enforcement of the Law in Social Networks (the “Network Enforcement Act”). The Network Enforcement Act requires all German telemedia service providers, as well as non-German telemedia service providers that satisfy specific requirements, to publish semiannual reports and reply to complaints about unlawful content within a specific timeframe. It names the Federal Office of Justice as the administrative authority.

3.2

Self-regulatory and Co-regulatory Instruments

3.2.1. In contrast to legal instruments, which usually take the form of legislation and require administrative action, co-regulation and self-regulation also play a crucial role in content moderation.

3.2.2. Where ISPs and OSPs implement these instruments, they can take various forms to suit their particular circumstances.

3.2.3. Most of the co-regulation and self-regulation measures are discussed in the following paragraphs. They include setting policies, structuring interactions, and monitoring and evaluation. Practices include deleting or modifying content, blocking users, creating access or filtering rules, and imposing temporary bans.

3.2.4. All these measures make it possible for service providers to respond voluntarily and at their own discretion. Because here it is the platform that makes decisions about content, and not the courts, the decisions may be attacked as a form of private censorship. In practice, however, service providers frequently implement these measures in pursuit of

their business interests and to maintain the balance among the various consumers they want to serve.⁸⁶

3.2.5. In what follows we address several co-regulatory instruments, industry-level self-regulatory instruments, and company-level regulatory instruments.

3.2.5.1. Co-regulatory instruments:

3.2.5.1.1. In co-regulation, the responsibility for the drafting and enforcement of regulations is shared by the state, the regulated market, and, in many cases, by intermediaries that interact with the regulators and the regulatees.

3.2.5.1.2. Whereas the specific regulatory arrangements may vary as a function of the particular circumstances of the regulated material, the regulatory regime's cooperative techniques and legitimacy derive, at least in part, from public-private cooperation.

3.2.5.1.3. Joint definition of market-based agreements:

3.2.5.1.3.1. Market-level policies are a relatively new instrument, because they require some supranational or national legitimacy.

3.2.5.1.3.2. For example, in May 2016 the European Union signed an agreement with four of the most important OSPs—Facebook, Microsoft, Twitter, and Google (for YouTube)—on countering illegal hate speech online. The agreement allows OSPs to strengthen their cooperation with other platforms.⁸⁷

86 DAVID S. EVANS & RICHARD SCHMALENSSEE, *MATCHMAKERS: THE NEW ECONOMICS OF MULTISIDED PLATFORMS* (2016).

87 [Code of Conduct on Countering Illegal Hate Speech Online](#) (May 31, 2016).

3.2.5.1.3.3. The joint agreement defined a code of conduct, based on the conditional liability of the E-commerce Directive and Framework Decision 2008/913/JHA. It requires the removal of content within an appropriate timeframe following a valid notification.

3.2.5.1.3.4. OSPs are also required to have clear and effective procedures to review notifications, to vet most requests against their rules and community standards within 24 hours, and to decide to remove or disable access to content if necessary.

3.2.5.1.3.5. The code also requires platforms to educate their users and employees and raise their awareness, to draft procedures for users and trusted reporters to submit notices and flag content, and to increase their best-practice training of civil society organizations (CSOs) to counter hate speech and to promote better and more effective campaigns to counter hate speech.

3.2.5.1.3.6. By signing this agreement, the OSPs formally joined the efforts by the European Commission and EU member states to ensure that online platforms do not offer opportunities for the viral spread of illegal online hate speech.

3.2.5.1.3.7. Although other global, market-based mechanisms do exist,⁸⁸ market-based policies can also exist on the national level. For instance, the

88 See, e.g., the Global Internet Forum to Counter Terrorism (GIFCT). At the GIFCT, large OSPs work with smaller technology companies to share insights about terrorism trends.

ISPs in the United Kingdom established an industry association that enforces codes of conduct to prohibit hate speech.⁸⁹

3.2.5.2. Industry self-regulation policies

3.2.5.2.1. Self-regulatory policies work without direct government involvement. Although a single company, several companies, or the entire industry have initiated self-regulating policies, they usually exist in the shadow of public policies.

3.2.5.2.2. Self-regulation policies can take different forms, including industry self-regulation, company-level policies, community-wide standards, and community composition policies, which are policies drafted by community members. The next paragraphs address these different forms more broadly:

3.2.5.2.2.1. In self-regulation, the industry sets and enforces non-binding rules.

3.2.5.2.2.2. In markets where there are “soft laws” and codes of conduct, government agencies can change their reaction from supervising the industry’s actions to encouraging the industry to meet its objectives.

3.2.5.2.2.3. One form of industry self-regulation to combat hate crimes is technology-driven and calls for the application of a particular production or process

89 This is despite the British law absolving ISPs and digital service providers of liability for hate speech. See James Banks, *Regulating Hate Speech Online*, INTERNATIONAL REVIEW OF LAW, COMPUTERS & TECHNOLOGY 24:3 (2010), at 233.

technology. Companies may use these technologies, but they are not required to do so.

3.2.5.2.2.4. For example, the industry can develop and use databases so that companies can share information. In May 2016, the four IT giants—Facebook, Microsoft, Twitter, and Google (for YouTube)—announced a new mechanism for sharing digitally signed hashes of terrorist content and recruitment videos for terrorist organizations.⁹⁰ The shared hashes will represent content identified and marked on one platform and will enable other platforms—including other (smaller) firms that are not parties to the project—to delete questionable content even before they have identified it as problematic on their platforms. Because one company warns another company about the existence of illegal or problematic content, to some extent this warning replaces the notification by law-enforcement agencies that is part of the conditional liability model. According to the industry statement, although the shared information will include only “extreme” cases of terrorist content, which will most likely violate all companies’ policies, the companies will retain their discretion to decide whether the content in fact violates their policies.

90 This Hash Database is part of the broader Global Internet Forum to Counter Terrorism initiative (GIFCT), in which Facebook, Google (for YouTube), Microsoft, and Twitter joined together to develop technological solutions, conduct research, share knowledge, engage with smaller companies, and promote counter-speech. See: *Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism*, FACEBOOK NEWSROOM (June 26, 2017).

3.2.5.3. Company-level self-regulatory policies

3.2.5.3.1. Self-regulatory policies at the company level can also influence how service providers moderate content on their platforms.

3.2.5.3.2. Two types of policies are of most relevance: contractual and organizational. We address them below.

3.2.5.3.3. **Contract-based mechanisms:** The contracts between OSPs and their customers state each company's expectations regarding the behavior of its customers in legal terms. Because the OSPs publish these policies and customers must agree to them in order to access the service, the OSPs have a legal basis for removing offensive content that violates their policies and for evaluating and punishing users' behavior.

3.2.5.3.3.1. These policies include community standards, user codes of conduct, and terms of service (TOS).

3.2.5.3.3.2. Unlike terms of service, which are contractual, community standards and codes of conduct are usually quasi-voluntary legal agreements that customers must accept. They make it possible for OSPs to regulate third parties and their users.⁹¹

3.2.5.3.3.3. By means of these statements and policies, ISPs and OSPs can delete content, disconnect users for a predefined time, or banish users who breach their contractual obligations.

⁹¹ In fact, even if the source of the content is located within the U.S., and thus enjoys broad First Amendment protection, service providers can remove content for violating their agreements.

3.2.5.3.3.4. For instance, the policies of OSPs like Facebook, Twitter, and YouTube define hate speech as unwanted behavior. This definition allows the companies to moderate the content on their platforms and avoid provoking controversy.

3.2.5.3.3.5. Whereas YouTube's TOS state that the platform is not liable for offensive content, its Community Guidelines require users to "respect the YouTube community" and warn users not to abuse the site.

3.2.5.3.3.6. In a later section, the Community Guidelines discuss the tension between free speech and hate speech and their regulation.⁹² Similarly, Twitter's TOS and Facebook's Terms of Service (previously called the "Statement of Rights and Responsibilities") disclaim the platform's liability,⁹³ while the "Twitter Rules" and Facebook's "Community Standards" discuss platform norms.⁹⁴

92 "Our products are platforms for free expression but we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics." YouTube, [Community Guidelines](#). See also Sellars, *supra* note 27.

93 As part of its contractual conditions, Facebook's Statement of Rights and Responsibilities references a set of community standards. Users may not use Facebook products to do or share anything that violates Facebook Community Standards See Facebook, [Community Standards](#).

94 Within a section of those rules entitled "abusive behavior," Twitter specifically prohibits "hateful conduct," defined as "promot[ing] violence against or directly attack[ing] or threaten[ing] other people on the basis of race, ethnicity, national origin, sexual

3.2.5.3.3.7. Platforms may also send a variety of messages or communicate through their user interface. Here the platform can provide users with examples of acceptable and unacceptable conduct. The idea is that notifying users of community guidelines will deter prohibited behaviors.

3.2.5.3.4. **Organizational policies:** Some companies adopt self-regulation instruments to address a policy problem.

3.2.5.3.4.1. Organizational policies are internal to the company and address how it responds to a legal or contractual breach.

3.2.5.3.4.2. By means of these policies, companies revise their structure and procedures to reduce the existence of bias or hate crimes.

3.2.5.3.4.3. Companies may modify their organizational makeup and policies and devise procedures to deal with unwelcome social phenomena.⁹⁵

orientation, gender, gender identity, religious affiliation, age, disability, or disease." Twitter also makes clear that it does not allow accounts "whose primary purpose is inciting harm towards others on the basis of these categories."

Facebook, on the other hand, identifies hate speech subject to removal from the platform as "content that directly attacks people based on their race; ethnicity; national origin; religious affiliation; sexual orientation; sex, gender, or gender identity; or serious disabilities or diseases." Beyond this, Facebook bans "[o]rganizations and people dedicated to promoting hatred against these protected groups." In contrast, Facebook considers sharing of hate speech "innocent" when said sharing contains "someone else's hate speech for the purpose of raising awareness or educating others about that hate speech." See also Sellars, *supra* note 27.

⁹⁵ On setting policies in the context of discrimination, See Levi & Barocas, *supra* note 55. According to Levi and Barocas, companies

3.2.5.3.4.4. In 2017, for example, Facebook announced that it was hiring an additional 3,000 content reviewers, for a total of 7,500. These reviewers supplement the policy analysis teams and policy directors it already employs worldwide.⁹⁶ The presence of moderators all over the world affords diversity in decisions about content moderation. News platforms and corporations, by contrast, usually have editors who must approve content, and in some cases also comments, before they are uploaded to the website.

3.2.5.3.4.5. Companies can also educate and train workers or create internal codes of best practices. Companies like Facebook and Google (for YouTube) already have such organizational policies installed. For instance, Facebook's abuse standards operations manual (2012) instructed content moderators to flag nine different forms of hate content. It stated that humor overrules hate speech unless slur words are present or the humor is not obvious.⁹⁷ It also

fighting discrimination will increase the representation of underrepresented groups within their engineering teams or invest personnel and other resources to eliminate bias.

⁹⁶ Kathleen Chaykowski, *Facebook Is Hiring 3,000 Moderators in Push to Curb Violent Videos*, FORBES (May 3, 2017).

⁹⁷ For instance, the 2012 abuse standards included: (1) slurs or racial comments of any kind; (2) attacks based on a protected category; (3) hate symbols, either out of context or in the context of hate phrases or support of hate groups; (4) shows of support for organizations and people primarily known for violence; (5) symbols primarily known for hate and violence, unless comments are clearly against them; (6) "versus photos" comparing two people (or an

mentioned political speech. In the manual, Facebook listed the categories that are subject to filtering and content moderation, including race, ethnicity, national origin, religion, sex, gender identity, sexual orientation, disability, and any serious disease.⁹⁸

3.2.5.3.4.6. Facebook's newer guidelines differentiate between problematic content that leads to automatic removal and content that is not problematic. For example, its hate-speech policies call for deleting content that includes curses, slurs, and calls for violence against "protected categories" such as "white men" when both the group and the subset are protected. On the other hand, it allows users more leeway when they write about "subsets" of protected categories, such as "black children" or "female drivers" that have attributes of groups that are not protected (children and drivers).⁹⁹

3.2.5.4. Algorithm-based instruments:

3.2.5.4.1. Companies can also decide to implement smart algorithms as a company-level self-regulatory measure.¹⁰⁰

animal and a person that resembles that animal) side by side; and (7) Photoshopped images showing the subject in a negative light.

98 See Sellars, *supra* note 27.

99 Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children*, ProPublica, June 28, 2017. Facebook has since changed this policy; see Josh Constine, *Facebook Reveals 25 pages of Takedown Rules for Hate Speech and More*, TECHCRUNCH (April 24, 2018).

100 Based on a lecture by Dr. Omri Abend, The Hebrew University of Jerusalem, at the workshop *Combating Online Hate Speech*, hosted by the Israel Democracy Institute on November 7, 2018.

In fact, artificial intelligence can execute natural language processing (NLP) techniques to process large amounts of (text) data and draw insights otherwise impossible to achieve.¹⁰¹

3.2.5.4.2. For instance, NLP is used to help with common search terms (as in Google Auto-Complete) and to provide services such as digital agents that can communicate in a pseudo-human manner (Alexa, Siri, Google Duplex). For advertisers, NLP means both the capability to compile terms obviously related to their brand but also to reach new consumers by capitalizing on uncommon terms.

3.2.5.4.3. In addition to improving advertising revenue and services, OSPs can use NLP to correct errors and spelling mistakes, retrieve information, and identify hate speech using text classification. The main paradigm for classifying text is called “supervised learning.” The first step in supervised learning is the labeling of data, usually by human experts who decide whether a text contains hate speech or not. These annotated texts are then fed into a predictive model that tries to learn and generalize. The last step is to apply what the model learned to new data that is not labeled and to make a prediction.

3.2.5.4.4. There are different features the system can be coded to pay attention to. These features are types of information with computable characteristics that we hypothesize to be related to the prediction. The features, or

101 Academic literature on machine detection of hate speech can be found in more common languages such as English, German, and Dutch. But the technology is fairly simple and well understood and can also be applied in other languages.

their combination, are later used to make decisions. There are different features we can use:

3.2.5.4.4.1. **Wordlists:** One possibility is to list the words and expressions that OSPs identify as prohibited. The words and expressions can be general or may be specific to a country or a group. There are several limitations to using wordlists: First, words used in posts are context-sensitive. Second, as languages keep changing and updating, wordlists have limited coverage. Creating the lists is laborious, but the lists also have to be updated constantly.

3.2.5.4.4.2. **Bag-of-words:** With this technique, humans encode all the words in the text, and sometimes their combinations (pairs or triplets), and let the system decide which of them are indicative or contraindicative, and whether the text includes hate speech or not. The bag-of-words technique offers some additional benefits over the previous techniques. It is more flexible, thanks to the possibility of adding and annotating new training data and helping the system adapt. The bag-of-words technique is also fairly transparent, because it is possible to tell which words actually triggered the system. There is a limitation, however: because it cannot be generalized across words, a word that did not appear in the training data will not be marked as problematic.

3.2.5.4.4.3. **Deep-learning technologies:** Deep-learning technologies are used to find words that share a distribution pattern and then conjecture that they are somehow related. This method has been very useful in NLP and has registered considerable

achievements. However, deep-learning technologies add noise and can trigger alerts in cases that are not really problematic as well as make the results more opaque. Mainly, the word embeddings and the technology used to generalize across words make it difficult to understand exactly what it is doing. In short, deep learning is more effective but also less transparent.

3.2.5.4.4.4. Character embeddings: Often, words are misspelled – sometimes accidentally (omission of vowels or letters) and sometimes deliberately (e.g., use of \$ instead of S or of the digit 1 instead of the letter l). Users who want to post hate speech may misspell words in order to bypass detection. Character embedding tries to adapt to these misspellings and deploys techniques for understanding the meaning of characters and not only of complete words.

3.2.5.4.5. Context sensitivity: In attempts to find out what hate speech is (and generally in attempts to classify text), context plays a key role. For instance, although the bag-of-words approach pays attention to the words being used, it is indifferent to the order in which the words appear in the linguistic and discourse structures. Some technologies try to tackle this problem, but they are more language-dependent:

3.2.5.4.5.1. Sentiment analysis: Sentiment analysis is a method that seeks to determine whether a given text expresses a positive or negative sentiment. If the text contains high-intensity negative sentiment, a warning that something problematic might be going on there can be triggered. However, while it is becoming easier to detect strong sentiment or specific words, technology is still limited in its

capacity to identify more complex categories such as sarcasm or newsworthiness.

3.2.5.4.5.2. **Linguistic structure:** Understanding the linguistic structure of a language can help pin down differences between similar texts.¹⁰²

In some cases, however, linguistic analysis might be harder to deploy—for example, when a text that might not include hate speech or emphatic language turns out to correlate with problematic language. Other cases relate to pejorative terms and require more precise language analysis; e.g., “the gays” and “the illegals” are more offensive than “gay people” or “people who have entered the country illegally.”

3.2.5.4.6. **New frontiers:** There are some emerging NLP technologies that have not yet been tested:

3.2.5.4.6.1. **Multimodal information:** Multimodal information analysis goes beyond text to include images, audio, and video, which might help understand speech on social media, achieve better predictions, and be more accurate at flagging problematic content.

3.2.5.4.6.2. **Structure-based approaches:** This technique analyzes speech by recognizing implicit structures in the discourse; e.g., what role is taken

102 For instance, we can think of this example: “Jews are lower-class pigs” and “Probably no animal is disgusting to Jewish sensitivities as the pig.” Both sentences contain “Jews” or “Jewish” and “pig,” yet knowing a bit more about the linguistic structure of English can aid in identifying that only the first sentence should be considered under a hate-speech takedown policy.

by participants and therefore whether each one is a bully, a victim, a defender, or a bystander.

3.2.5.4.6.3. **Inference:** Inference remains a difficult task for machines to perform; this applies notably to sarcasm, mockery, and implicit abusive language. In these cases, a text might be acceptable in some circumstances and offensive in others. But it might be hard to determine the circumstances of the particular case. Examples are “Kermit called and wants his voice back” to mock someone’s voice or “Put on a wig and lipstick and be who you really are” to mock a person’s sexuality or gender identity.

3.2.5.4.6.4. **Identifying intent:** Another frontier is the identification of intent – intent to harm, to cause additional harm beyond the speech itself, or to incite to socially undesirable actions. Intent is frequently implied and machines may not be able to identify it.

3.2.5.4.7. Given these capabilities and limits of algorithm-based NLP mechanisms, at present they can be used to automatically identify, filter, or flag harmful or illegal content in the following ways:

3.2.5.4.7.1. *Automatic filtering* replaces human decision-making for the OSP. Both flagging and removal of content are automated.

3.2.5.4.7.2. *Automatic flagging* replaces decisions by users and trusted flaggers. Here, unlike in automatic filtering, a human must still decide to remove the problematic content.

3.2.5.4.7.3. *Automatic approval of legitimate content:* In both automatic filtering and flagging,

the algorithm can scan and automatically approve content.

3.2.5.4.7.4. *Automatic approval of questionable content*: After a service provider has viewed questionable content, it can automate the decision. For example, if the OSP has decided to retain some flagged content on its service, it can automatically notice future flaggers of this decision. If the OSP chooses to take down the content, it can automatically remove similar content.

3.2.5.4.8. The *New York Times* has partnered with Alphabet's Jigsaw to develop machine-learning tools to moderate the *Times's* online comments section. This algorithm-based mechanism, appropriately called "Moderator," was trained on more than 16 million previously moderated *Times* comments. "Moderator" automatically prioritizes comments that are likely to require review or removal and thus substantially increases the volume of allowed comments.¹⁰³

3.2.5.5. Structuring user interactions:

3.2.5.5.1. During the process of platform design, every OSP also considers how to structure interactions among users. In some cases, this decision is based on a prior decision about the composition of the community; that is, whether the platform is for all audiences or specifically for a particular group.

3.2.5.5.2. With regard to the structuring of interactions, OSPs, through their platform's user interface (UI), can control

103 Bassey Etim, *The Times Sharply Increases Articles Open for Comments, Using Google's Technology*, NEW YORK TIMES, June 13, 2017.

what users learn about other users' characteristics, as well as what information and content will flow between users.¹⁰⁴

3.2.5.5.3. When an OSP decides on a user interface that supports interaction, it exercises control over the types of information that other users can access.

3.2.5.5.4. By means of their platform, OSPs can encourage or require the disclosure of information, withhold user information and content, structure the input of user information, or link user information to external sources of information.¹⁰⁵

3.2.5.5.5. A simple example of the structuring of interactions involves users' control of their profile display (such as an extended profile to "friends" and a limited profile to others). Companies like Facebook can require real-name user profiles, while Twitter can allow users to employ generic names or hashtags. This decision can have consequences for users' ability to choose usernames or hashtags that are themselves hate crimes or offensive to a specific group.

3.2.5.5.6. Another essential feature of interactions is whether the connection between two users is one-way (e.g., Twitter or YouTube) or bidirectional (Facebook and LinkedIn):

3.2.5.5.6.1. Bidirectional connections require both users to approve the "friendship" before the platform creates a link for information and content-sharing between them. For instance, the connection

104 On discrimination, see Levi & Barocas, *supra* note 55. On privacy regulation, see Rotem Medzini, *Prometheus Bound: A Historical Content Analysis of Information Regulation in Facebook*, *JOURNAL OF HIGH TECHNOLOGY LAW* XVI: 1.5, at 195.

105 For further elaboration on the moderation of bias on social media, see Levi & Barocas, *supra* note 55.

between Facebook friends is bidirectional, which means that users cannot post content on another user's wall without the latter's consent. But if the two users are Facebook friends, posting or tagging users can be much easier.

3.2.5.5.6.2. One-way connections enable one user to "follow" and receive updates from another user. This is the case on Twitter and the meaning of following a user or page on Facebook. Even when two users follow each other in this manner, they are not in a bidirectional connection; at any time one of them can decide to stop following the other without consequences to the connection in the other direction. Only blocking the other user will sever both connections.

3.2.5.5.7. Companies can also structure their platforms' user interfaces so that users can influence the rank and importance of content posted by other users. "Liking" or reposting content is one such form of control. On the individual level, liking or reposting notifies a user's friends of content the user deems exciting or important. On the collective level, liking or reposting makes a post go viral. User interfaces can also allow users to change the rank of the content that specific users will receive. On Facebook, for instance, the platform enables users to tell Facebook which friends should receive privileged access to the wall or whose posts should receive priority on the newsfeed.¹⁰⁶

106 Platforms such as Facebook sometimes enable users to have stronger control over visible content, including limiting their friends' option to post content on their wall or lowering their friends' posting on their news feed.

3.2.5.5.8. Automatic content selection by means of a smart algorithm is another way in which companies structure interactions among users. For many OSPs, the ability to suggest up-to-date and relevant content to users is an important element of their business model and need to remain relevant. For Amazon, this means the ability to recommend to users what other shoppers have looked at or bought along with a specific product. For Facebook, it is the ability to present relevant and popular content posted or tagged as interesting by friends at the top of the news feed.

3.2.5.5.9. For Google, unfiltered videos on YouTube may lead to a suggestion of other unfiltered content viewers might want to watch next. In order to combat negative forms of content bubbles, such as those that contain a collection of white nationalist videos, OSPs can implement a video-selection algorithm to safeguard and sanitize all or parts of their service or execute counter-speech initiatives.¹⁰⁷ In this way, OSPs can decide who will be the audience of hate speech and determine whether or not it will go viral. In the wake of public comments, for instance, YouTube promised to implement stricter standards on extremist content. According to Susan Wojcicki, CEO of YouTube, in 2017 YouTube tightened its policies about what content can appear on the platform or earn revenue for creators. Content that violates YouTube's policies is to be removed quickly, while content that does not necessarily violate specific rules

107 Such examples include YouTube's Creator for Change, Jigsaw's Redirect Method, Facebook's P2P and OCCI, and Twitter's NGO training program. See *Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism* (June 26, 2017).

can be limited through warnings, restriction of the ability for it to be monetized with advertising, and a ban on posting it as recommendations, endorsements, and comments.¹⁰⁸ This makes it harder for policy-violating content to surface or remain on YouTube and ensures creators and advertisers of stability for their brand names and revenue.¹⁰⁹

3.2.5.6. User interfaces and flagging mechanisms:

3.2.5.6.1. Companies can also provide users with mechanisms to limit unwanted interactions (with or without relevance to hate crime). Privacy settings, for instance, enable users to designate who can have access to their content and private data. When a platform such as Facebook promotes the freer flow of information to increase content virality, it also modifies users' privacy settings and makes users more approachable by content they may prefer not to see.¹¹⁰ Providing users with the right and facility to adjust their privacy settings allows them to decide who sees the content they are sharing as well as which content they prefer not to see.

3.2.5.6.2. OSPs can provide users with ways to flag content as seemingly offensive or socially deviant and thus a candidate for moderation. However, the data the platform can request as part of the flagging procedure may vary from report to report.

108 Daisuke Wakabayashi, *YouTube Sets New Policies to Curb Extremist Videos*, NEW YORK TIMES, June 18, 2017.

109 Susan Wojcicki, *Expanding Our Work against Abuse of Our Platform*, YouTube's Blog (December 4, 2017).

110 Medzini, *supra* note 104.

3.2.5.6.3. By means of report systems, an OSP can ask users to provide granular information on the case, so that it can obtain more details about the reported content.¹¹¹ At the same time, it is important to note that imposing too many requirements and demanding too much information as part of the report can make it cumbersome and discourage users from reporting problematic content or events.

3.2.5.6.4. Although flagging mechanisms are not always easy to implement and can be used for abuse—for example, to falsely report hate crimes as a way to have legitimate content that the reporter does not agree with removed—these mechanisms are critical for content moderation. YouTube, for example, permits users to hide content they find inappropriate without having to notify YouTube of its existence and whether or not the content in fact contains hate speech.

3.2.5.6.5. OSPs can also decide to keep previously reported content online. In these situations, a repeated flagging of the same content may lead the OSP to decide to take down the content following a secondary review or notice to users of the previous decision to keep the content online.¹¹² Also, as reported by Facebook, previously flagged content that the platform has decided to keep online can be automated through the platform's algorithms, thus saving the company the need to make the same decision until the facts or the content change.

111 On implementation in reporting discrimination, see Levi & Barocas, *supra* note 55.

112 This process can be automated. See The Berkman Klein Center for Internet & Society, *The Line Between Hate and Debate on Facebook*, The Berkman Klein Center for Internet & Society (September 22, 2017).

3.2.5.7. With these front-end mechanisms, companies can learn to take cues from their users, moderate content, and adapt their back-end procedures. All these “small” decisions can influence whether two users are aware of one another, and consequently whether hate speech passes between users.

3.3 Information-Based Instruments

3.3.1. A third method for challenging hate speech employs information-based instruments—the use of information as a resource to alter behavior.

3.3.2. The leading promoters of information-based instruments are civil society and the media. But law-enforcement agencies, teachers, and OSPs can also provide information and educate.

3.3.3. There are a number of information-based mechanisms: public monitoring, public advocacy, research advocacy, agenda settings, advocacy journalism, the flagging of hate crimes, education, and cyber-literacy. We address each of these mechanisms below:

3.3.3.1. Public monitoring:

3.3.3.1.1. Civil society, and sometimes other actors as well, can track hate speech and xenophobia and provide information on the extent to which they are present on an online platform.¹¹³

3.3.3.1.2. Another valuable source for information about hate crimes is the OSP’s annual transparency report on takedown requests by law-enforcement agencies.

113 For broader examples, see chapter 1.

3.3.3.2. Public advocacy:

3.3.3.2.1. Public policy advocacy can take the form of the production of guidelines and best practices for responding to hate speech online.

3.3.3.2.2. By writing guidelines and codes of best practices, civil society can teach policymakers and OSPs how to improve the legal and self-regulatory responses to online hate speech.

3.3.3.2.3. For example, civil society can produce brochures on issues such as net-neutrality or on how the internet works.¹¹⁴ Civil society can also develop best practices for OSP responses to online hate speech.

3.3.3.2.4. In these codes of best practices, civil society can recommend that law-enforcement agencies and OSPs take reports of online hate speech seriously, explain to users the platform's approach to resolving online hate speech reports promptly, and offer user-friendly mechanisms for reporting online hate speech.

3.3.3.3. Research advocacy:

3.3.3.3.1. Advocacy can take the form of research, in the belief that research is the first step in exposing online threats.¹¹⁵

114 In the European context, see [European Digital Rights](#) (online).

115 On civil society organizations in the privacy policy debate and counter-surveillance advocacy, see COLIN J. BENNETT, *THE PRIVACY ADVOCATES: RESISTING THE SPREAD OF SURVEILLANCE* (2008).

3.3.3.3.2. Although research advocacy usually derives from socially aware academics, civil society can also develop databases that contain research-based content about hate speech.

3.3.3.3.3. For instance, the Anti-Defamation League maintains a database of different OSPs' hate-speech policies¹¹⁶ and publishes a report on the increase in hate crimes. The Pew Research Center issues quantitative reports about current online phenomena, including hate speech. The Electronic Frontier Foundation publishes annual transparency reports on OSPs' sharing of information with state actors. EPIC (the Electronic Privacy Information Center) tracks advocacy actions and follows changes in the information practices of OSPs.

3.3.3.4. Agenda-setting and advocacy journalism:

3.3.3.4.1. Civil society and the media can educate policymakers and the public at large and ensure that the problem of hate speech never falls off the public agenda.

3.3.3.4.2. For instance, the media can make the public and policymakers aware of the extent of the phenomenon and report new challenges created by new information and communication technologies. Media organizations can headline the reports issued by civil society organizations, thus setting the public agenda.

116 [ADL Cyber-Safety Action Guide, ADL \(online\)](#).

3.3.3.5. Flagging hate crimes:

3.3.3.5.1. Civil society organizations can act as trusted flaggers and help ISPs, OSPs, and law-enforcement agencies identify content as hate speech and trigger automated flagging mechanisms.

3.3.3.6. Education and cyber-literacy:

3.3.3.6.1. Civil society, as well as service providers and educators, can educate citizens about correct and safe use of the internet and online platforms. Platforms can teach about different practices that implement the instruments mentioned above.

3.3.3.6.2. Educational and awareness-raising materials can teach citizens, and especially children, how to identify hate crimes, how not to create hate speech, how to notify law-enforcement agencies and companies about hate speech, and how to reduce its impact.

3.3.3.6.3. Education does not deal with the instigators but instead aims to mitigate the effects of hate speech after it occurs.

Chapter 4

The Proposal: A Co-regulation Model with Common Criteria for Defining Hate Speech

In this chapter, we offer a model for dealing with hate speech on social-media platforms. The model is co-regulatory and includes two key aspects: common criteria for identifying hate speech, and a detailed co-regulatory application procedure. We discuss each of these aspects below. In the next chapter we describe what led us to select this model in preference to the others presented above.

First, we offer common criteria for identifying hate speech. Here we are building on the examples we presented in Chapter 1 and on the work of Andrew Sellars.¹¹⁷ We crafted our criteria in the form of continua to enable OSPs to visualize their chosen policy logic, on the range from a more conservative to a more lenient content policy.

Second, the model includes a co-regulatory mechanism for implementation. We propose a design in which OSPs and law-enforcement agencies share responsibility for moderating hate speech: OSPs create procedures to moderate content, while law-enforcement agencies notify them of problematic content.

To clarify, we do not suggest a pre-upload content-moderation model and do not intend to get involved in the current and common business model of the OSPs.¹¹⁸ Because we are aware that OSPs provide forum, groups,

¹¹⁷ Sellars, *supra* note 27.

¹¹⁸ Recently, upload regulation of content was mentioned in regard to Article 13 of a proposed directive on copyright in the Digital Single Market, which would require information-society service providers (an EU term that includes OSPs) to take measures to ensure the functioning of their agreements with rights-holders for the use of their works or to prevent the availability on their services of works and other subject matter identified by rights-

and pages managers with mechanisms for moderating upload content, we suggest that in such cases managers should bear liability for content published on their page, just like private individuals on their private pages.

Chapter 4(a)

Common-Criteria Definition of Hate Speech

The first part of our model is based on common criteria to identify hate speech. We are basing these criteria on the comparison in Chapter 1 and on the work of Andrew Sellars, who identifies eight factors that categorize speech as hate speech or as speech that might lead to hate-related offenses.¹¹⁹ We use Sellars' criteria because his definitions reflect what most countries and the major platforms would define as "hate speech," including actionable hate speech in the United States. However, we do not attempt to define hate speech as a legal normative or positive criterion, but rather leave the decision on the exact policy to the OSPs. Our common criteria break the broad definition of hate speech into smaller definitions scaled on several continua that range from a more conservative to a more

holders. According to Article 13, these measures including the use of effective content-recognition technologies and should be appropriate and proportionate. According to a resolution passed by the European Parliament, online content-sharing services (another EU term that includes OSPs), as an act of communication to the public, shall conclude fair and appropriate licensing agreements with rights-holders. Only in the absence of a licensing agreement must an online content-sharing service provider take appropriate and proportionate measures leading to the non-availability of works on those services. See [Amendment 78, Report on the Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market \(COM\(2016\)0593 – C8-0383/2016 – 2016/0280\(COD\)\)](#) (June 29, 2018).

¹¹⁹ Sellars, *supra* note 27.

lenient content policy. This visualization in turn enables the OSPs to better understand where they choose to place themselves on each continuum.¹²⁰

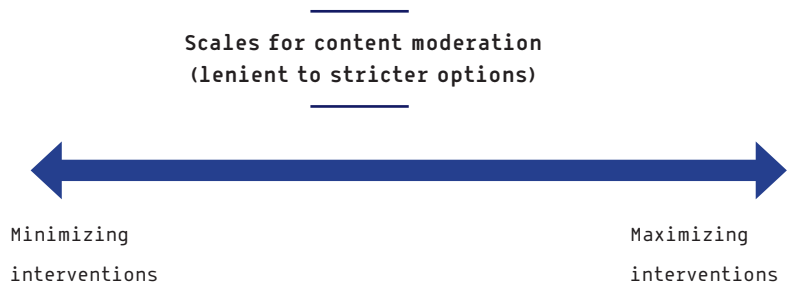
Our analysis of the common criteria fits in with our decision not to leave the criteria as definitions but instead to create a continuum of scalable options for each of them. In this way, our common criteria provide a decision-making mechanism for OSPs for the implementation of each criterion and whether it should be implemented in a lenient or stricter manner. Using these continua, OSPs can more easily define a uniform policy on where they want to stand on moderating hate speech without the need to pick and choose between vague policies that might or might not be relevant to content containing hate speech.

In addition, our analysis makes it possible for every OSP that develops and runs a social-network platform to define its ethical position—its overall policy on combating hate-speech and its position on each criterion. If some managerial decision does not coincide with the social network's economic model or creates political controversy, the company's executives can move along the continuum and choose another combination.¹²¹

We position each criterion along five continua. Each continuum supports a choice between the two poles: on the left side, more lenient options that enable less intervention in freedom of expression; on the right side, stricter options that lead to the deletion of more content. In some countries, of course, the government implements a stricter content-regulation regime and OSPs must choose between complying with the law or not providing their services in that country, for instance by means of geo-blocking.

120 For an example of the application of our model to Twitter's counter-hate-speech policies, see Appendix C.

121 ROBERT A. DAHL & CHARLES E. LINDBLOM, *POLITICS, ECONOMICS, AND WELFARE* (1953); Michael Howlett, *Policy Instruments, Policy Styles, and Policy Implementation: National Approaches to Theories of Instrument Choice*, 19 *POLICY STUD. J.* 1 (1991).



At the same time, given that hate speech, and sometimes specific content, may be illegal in some countries but not in others, OSPs need to deal with two issues. The first is what to do with countries without content limitations. This can lead the OSP to decide on transnational coverage or to geo-block content to specific countries that impose content limitations while leaving the content available to users in other counties. Second, the OSP must decide whether and how to harmonize content moderation in all countries that do regulate content. Such decisions can obviate geo-blocking for each particular country. The following paragraphs provide details of our scalable common criteria.

4.1. Common Criteria

(1) The speech targets a group or an individual as a member of a group: The most basic criterion for recognizing hate speech is that the speech either targets a group or targets an individual as a member of a group. This criterion distinguishes “hate speech” from other forms of harmful speech, such as defamation, bullying, or personal threats. Groups in this context may include minorities, historically oppressed and traditionally disadvantaged groups, or actionable groups, as described below:

Protected groups



**Racial, ethnic
and religious
groups**

Antisemitism,
Islamophobia,
African-American
hatred

**Other protected
groups**

Gender, sexual
orientation,
gender identity,
physical
disabilities,
serious diseases,
Holocaust
survivors

**Political, social
or professional
groups**

Party membership,
lobbying group,
ideology,
feminists, union
members, veterans

a. The most conservative definition of protected groups lists *race, ethnicity, and religion* as grounds for protection. These classifications directly link racism with the prohibition on discriminating against or speaking hatefully about a group or a member of a group. For instance, the definition of antisemitism promulgated by the International Holocaust Remembrance Alliance (IHRA) includes rhetorical and physical manifestations that are directed toward “Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities.”¹²²

122 The International Holocaust Remembrance Alliance (IHRA), ["Working Definition of Antisemitism."](#)

b. Several definitions protect people against hateful speech or discrimination based on membership in a protected group. While countries may protect people from discrimination, harassment, or hateful speech based on categories like sexual orientation, gender identity, or disability, these classifications are less directly linked to hate speech than is racism.

c. The most lenient definition protects voluntary groups. These may include political associations (e.g., political parties, lobbies, and ideologies such as Zionism), social cause and lobby groups (e.g., Planned Parenthood, the National Organization for Women Foundation (“NOW Foundation”), Black Lives Matter, trade unions, or AIPAC), or professional groups (e.g., US Army Veterans, the American Medical Association, the American Bar Association). As in the previous definition, some countries protect groups of this type against discrimination, harassment, or hate speech.

The decision to protect a group is usually based on global conventions as well as historical and cultural contexts. In addition, some groups are easier to define than others, and the definition can change depending on the OSP’s consumer public. This is why we do not offer a closed list of protected groups and leave the definition of protected groups to the companies’ discretion.¹²³

(2) The speech expresses hatred: The second criterion for identifying hate speech is whether the speech conveys hatred. Unlike the previous criterion, which refers to which groups are protected, this factor is usually open to national or legal interpretation. Additionally, rather than a continuum

¹²³ For instance, Facebook does not protect countries (Ireland, Britain, or the United States), political affiliation (Republicans or Democrats), people’s appearance (blond/brunette, short/tall, fat/thin), or social class (rich/poor). But it does have a quasi-protected category for migrants. See *The Facebook Files: Hate Speech and Anti-migrant Posts*, THE GUARDIAN, May 24, 2017.

that runs from limited hatred to more extreme statements, our proposed continuum reflects the decision about how the existence of hate speech is identified. Thus, OSPs' hate-speech guidelines must include procedures for identifying content that expresses hatred. Here too we offer several policy implementation options:

A closed list of definitions or symbols that represent hate speech typifies a policy that is more lenient, because it permits more content online. Policies that look at the content with regard to context (e.g., "some of my best friends are Jewish" or "Jews are very good with money"), newsworthiness, or legitimacy are more restrictive and can lead to more extensive removal of content. This is a "context-based approach."

—————
Definitions of expressions of hatred
(from closed list to context-based)
 —————



**Closed list of
definitions**

List of banned
expressions,
symbols, or
images

**Mixed
approach**

Bag-of-words,
conjunctions,
linguistic
structures

**Context-based
approach**

Satire,
historically
significant events,
newsworthiness

a. A *closed list* of definitions or symbols means there are predefined terms that may not be used.¹²⁴ Only if a term from the list is used should the content be taken down.

i. This approach is used in the United States, on First Amendment grounds.¹²⁵ On the one hand, a closed list provides certainty and is easier to enforce by means of algorithms.¹²⁶ On the other hand, closed lists are open to politicization; sometimes the terms that are left off the list are deemed socially acceptable even if offensive or harmful.¹²⁷

ii. *Symbols*, specifically, are graphical or textual representations that carry social messages, such as the swastika or the name “Hitler.”¹²⁸ This approach widens the list of terms to be taken down to non-textual representations as well as terms and expressions that employ socially offensive symbols.

b. A mixed approach: A *mixed approach* builds on the concept of NLP and supervised learning to label data, usually by relying on human experts to annotate data. The experts decide whether a text contains hate speech and define the words the algorithm needs to look for. The annotated texts

124 One such list is the Wikipedia list of [ethnic](#) or [religious](#) slurs. While this list was created by the Wikipedia community, other lists could be created through collaboration among OSPs, by civil society organizations, or through cooperation between OSPs and civil society (as we recommend in §4.5.9).

125 *United States v. Stevens*, 559 U.S. 460, 469 (2010).

126 For instance, Twitter has a closed list of behaviors it does not tolerate, including mass murder, violent events, and specific forms of violence in which groups have been the primary targets or victims.

127 Twitter, however, also deals with complexity by deleting groups whose “primary purpose” is inciting harm.

128 Facebook’s internal content guidelines place strong emphasis on symbols such as the swastika and on references to key figures notorious for hatred. See Appendix A.

can then be fed into predictive models that try to learn and generalize. Following this step, the models can then be applied to new data that is not labeled in order to make predictions on new texts. Different features of a mixed approach include “bag-of-words,” deep-learning technologies, and linguistic structures.¹²⁹

c. Context-based approach: A *context-based approach* examines the content within its context, given that even speech that expresses hatred may have some redeeming features,¹³⁰ such as satire or newsworthiness.

i. The idea here is that unlike closed lists, which do not recognize any legitimate use, the question of whether the content has some redeeming feature widens the range of acceptable content and relaxes the closed list approach. For example, Canada exempts certain types of speech, including speech that expresses “good faith” on a religious subject, speech that is true, and speech made in the public interest.¹³¹

ii. The relevant context can include the group the speaker is addressing, the type of expression, the offensiveness of the content, and the groups the content reached. Several social platforms’ providers use context when deciding about flagged content:

1. The Facebook community standards page indicates that content that might otherwise violate its standards may be allowed sometimes, but only if Facebook feels it is

129 Under “algorithm-based instruments” (starting in 3.2.5.4) above we discussed the capabilities and limitations of natural language technologies for identifying hate speech. Our recommendation here is based on the analysis there.

130 For instance, while the International Holocaust Remembrance Alliance (IHRA) provides rhetorical and physical examples of possible manifestations of antisemitism, it mentions that the overall context also needs to be taken into account. See The International Holocaust Remembrance Alliance (IHRA), *Working Definition of Antisemitism*, *supra* note 122.

131 Canada Criminal Code §319(3).

significant or important to the public interest. The decision is made after weighing the public interest against the risk of real-world harm.¹³²

2. Google tells YouTube users that they should add context to their videos and add key details to explain their videos, especially where graphic content is involved. As an example, Google explains that relevant information can include a list of tips at the beginning of the video, a clear title, or a description stating, for instance, that the video contains or documents harmful content. Adding key details, according to Google, helps other users find and understand the user's content and helps the YouTube team review the video if it was flagged.¹³³

iii. One key factor for understanding context is whether the context makes a violent response plausible. OSPs can consider several factors:

1. The speaker's power and status
2. The audience's receptiveness
3. The history of violence in the area where the speech takes place
4. The social and political context
5. The size of the audience
6. Whether, given the circumstance, it will stir up racial hatred

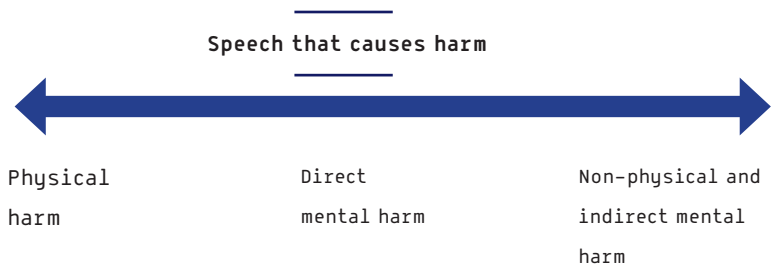
iv. Another option OSPs have is to use NLP to tackle some of the limitations of the wordlists and bag-of-words approaches

132 Facebook's community standards mention this balance under both safety and voice. See Facebook Community Standards, *supra* note 93.

133 See *The Importance of Context*, YOUTUBE HELP.

regarding linguistic and discourse structures. These approaches include sentiment analysis to identify negative sentiments and learning about the linguistic structure of the language to address differences between texts. Nevertheless, these technologies are still hard-pressed to identify sarcasm, understand the newsworthiness of the text, and handle less commonly used languages.¹³⁴

(3a) The speech could cause harm to an individual: This criterion addresses whether the content aims to cause additional harm beyond the speech itself. The criterion can be strict and include a call only for physical injury, or be more flexible and include a call for mental or indirect harm.



¹³⁴ For instance, following the discovery in 2018 that Facebook had not removed hate speech against the Rohingya and other Muslims in Myanmar, which led to a military crackdown and ethnic violence, it was revealed that Facebook had established a dedicated product, engineering, and policy team to specifically deal with content in Myanmar and increased its team of native Burmese speakers to 100 content reviewers (Facebook reported that it hired 99 of them—which means it lacked them until that time). Facebook also improved proactive detection of hate speech and misinformation in Myanmar and extended its use of AI to posts that contain graphic violence and comments. See Alex Warofka, *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar*, FACEBOOK NEWSROOM, November 5, 2018. See also Steve Stecklow, *Why Facebook Is Losing the War on Hate Speech in Myanmar*, August 15, 2018.

a. *Physical harm* means actual violence. Both the European Framework Decision and Twitter's terms of service bar content that aims to cause additional physical violence.¹³⁵

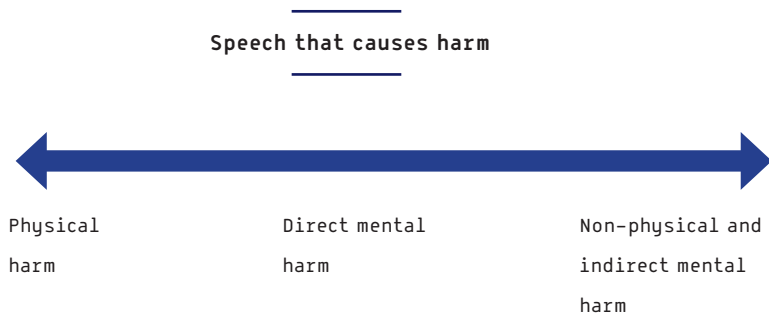
b. *Direct mental harm* can be a derivative of hate speech. It includes triggering fear and or frightening people about expressing their opinions.

c. *Non-physical and indirect mental harm* refers to hate speech that affects and influences the target's relationships with others, financial situation, performance at work, and social and personal life. It can include a refusal to hire or rent an apartment, which we do not see as falling into the category of physical or direct mental harm.

(3b) The speech could cause or provoke injury to a group: In addition to the possibility of injury to an individual, there is a similar continuum of hate speech aimed at a group. Statements in this category can lead over time to demonization, hostility towards the group, and legitimizing actions against the group.¹³⁶

135 Framework Decision 2008/913/JHA on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law (November 28, 2008); Twitter, *Twitter Rules*.

136 The IRHA's definition, for instance, includes targeting the State of Israel and a Jewish collective or "making mendacious, dehumanizing, demonizing, or stereotypical allegations about Jews as such or the power of Jews as collective" to control the media, economy, government, or other social institutions. See IHRA, *supra* note 122.



For instance, the United Kingdom investigates whether the circumstances of the speech are likely to stir up racial hatred.¹³⁷ In contrast, the Rabat Plan advises looking to the “social and political context,” the speaker’s status, and the size of the audience.¹³⁸

(4) The speaker intends harm: The importance of intent as a factor, whatever the difficulties of identifying it, derives from its close connection to the actual ability to cause harm.¹³⁹ The Rabat Plan identifies an intent to cause harm as an essential element of Article 20 of the ICCPR. The Facebook policy on harassment looks at both context and intent.¹⁴⁰ Google

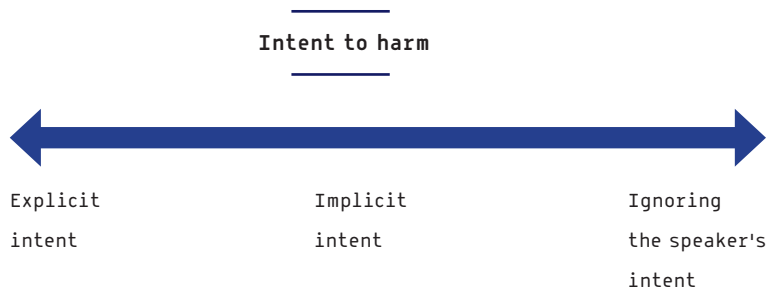
¹³⁷ Public Order Act 1986 §18(1).

¹³⁸ The Rabat Plan of Action ¶ 22.

¹³⁹ Sellars, *supra* note 27, at 28.

¹⁴⁰ Facebook defines harassment as sending messages that repeatedly contact large numbers of people with no prior solicitation and sending messages to any individuals that contain foul language aimed at an individual or group of individuals in the thread. Facebook does allow people to share and reshare posts if it is clear that the sharing was made to condemn or draw attention to harassment. See [Facebook Community Standards](#) (online). According to Facebook, while it looks at the context, it does try to discover the user’s intentions. See Richard Allan, VP EMEA Public Policy, *Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?* June 27, 2017. (hereinafter: Allan, *Hard Questions*)

formerly held that intent was an optional component of its assessment for YouTube,¹⁴¹ but now clarifies that if the user's action is repeated or coupled with malicious intent, there may be a stricter or longer reaction.¹⁴²



a. Explicit intent: The first option is to look only for clear and visible intent to cause physical or non-physical harm. For instance, Twitter targets conduct that promotes violence or directly attacks a group with the suggestion of underlying intent.¹⁴³ Canada looks for speech that willfully promotes hatred. For Facebook, content that appears to purposefully target private individuals with the intention of degrading or shaming them is subject to removal.

b. Implicit intent: Intent can also be implicit and have to be inferred from the context, the words used, or previous statements. Some NLP technologies such as sentiment analysis and linguistic structures try to tackle the problem of implicit intent. For instance, sentiment analysis can help determine if a text expresses positive

141 See Sellars, *supra* note 27, at 27.

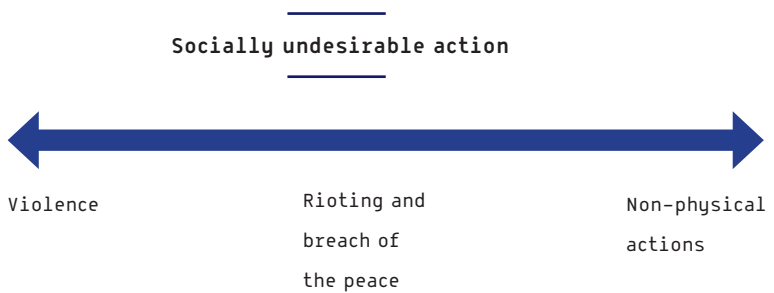
142 Normal responses include suspending ads, losing access to creator programs, and becoming ineligible for trending for a period of time. See Google, *Creator Influence on YouTube*.

143 See Sellars, *supra* note 27, at 27.

or negative sentiment. Multimodal information could also be used to go beyond text to learn from images, audio, and video.

c. The most lenient possibility does not consider the speaker's intent as a factor. In other words, any speech that falls under the other criteria mentioned in this chapter would be considered to be hate speech, whether or not the speaker had an intent to harm.

(5) The speech incites to socially undesirable action: This criterion addresses a requirement that the speech may incite other consequences. In the American context, the incitement must be imminent or almost inevitable.¹⁴⁴



a. *Violence*, such as murder or ethnic cleansing

b. *Rioting and breach of the peace*: Canadian law refers to speech that incites to a breach of the peace or to rioting,¹⁴⁵ as does the European Framework.¹⁴⁶

144 *Id.*

145 Canada Criminal Code §319.

146 Framework Decision 2008/913/JHA on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law (November 28, 2008).

c. *Non-physical action* includes content that calls on readers to humiliate individuals or to rally and protest outside homes and on the street (as in Charlottesville). Similarly, content can call on readers to distort the truth or spread disinformation and misinformation. Some legal definitions use a non-physical framework, such as intent to demean, humiliate, or incite hatred.¹⁴⁷ While Facebook looks at the context, it does try to discover the user's intentions.¹⁴⁸

Chapter 4(b) Procedures for Identifying Common Criteria and Content Moderation

4.2. Step 1: Implementing the Common Criteria for Identifying Hate Speech

4.2.1. Each OSP should institute company-level self-regulatory policies to implement the common criteria for identifying hate speech (chapter 4(a)). The internal procedures for reviewing notifications should be clear and effective.

The OSP's hate-speech policy must reflect decisions about the scales discussed in the previous chapter. The policy selected needs to include the specification that if content matches the criteria it is deemed to be manifestly illegal or undesirable on the platform and

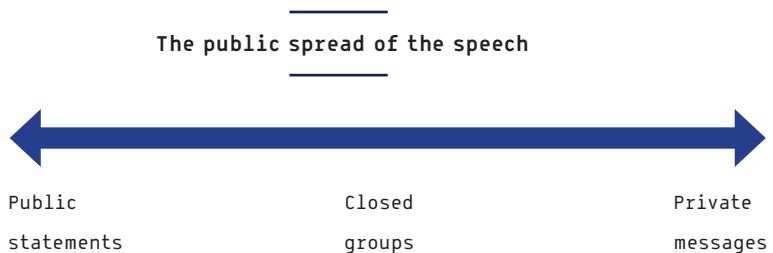
¹⁴⁷ For instance, the IRHA gives the following example of antisemitism: "Making mendacious, dehumanizing, demonizing, or stereotypical allegations about Jews as such or the power of Jews as a collective." The International Holocaust Remembrance Alliance (IHRA), *Working Definition of Antisemitism*, *supra* note 122.

¹⁴⁸ Allan, *Hard Questions*, *supra* note 140.

marked for immediate removal. At the same time, the policy needs to have grey areas where greater discretion is required.

4.2.2. The OSP's policy should reflect, among other things, the broader publication characteristics of the relevant platform and more dynamic rules based on the audience of the relevant post, which may or may not include hate speech. For instance, Facebook owns three platforms—Facebook, WhatsApp, and Instagram; each platform might have a different policy or all might have the same policy, but tweaked to its own preference.

4.2.3. The public spread of the speech: Content posted on social-media platforms can be visible to the general public (Twitter), to a closed group (Facebook), or to specific individuals (private messages on most platforms). Current laws (as in Canada and Australia¹⁴⁹) and proposals for legislation generally address only public statements. OSPs can moderate only content available to the public or content within closed groups as well. Moderating content within private messages is much less common.



149 Canada Criminal Code §319; Racial Discrimination Act 1975 §18C(2).

a. Public statements and open groups: OSPs can set the default of posts on their social media as public. For instance, most tweets are public and can be viewed and reshared by almost anyone, including those who are not Twitter users. Open groups are sectors of a social-media platform, such as pages, that any user can access or join without prior screening.

b. Closed groups: OSPs can decide that only the members of a closed group of users can access some content. Unlike open groups or pages, where users can decide whether or not to join the group, admission to a closed group usually requires the approval of the group administrator. The decision as to whether content is visible to everyone or to specific users only is usually left to the group administrator. Note that some closed groups are large enough to be considered public groups.

c. Private messages: Most social-media platforms permit users to send each other private messages that cannot be reshared. Some platforms allow users to forward private messages easily and only sometimes notify users that the message was forwarded.

4.2.4. As a function of their financial and technological abilities, OSPs should develop algorithm-based instruments for active monitoring and automatic flagging of questionable content, as defined by their policies regarding the common-criteria scale in chapter 4(a).

4.2.4.1. Content that violates the OSP's criteria should be flagged. Because such content violates the most stringent rules, it is important to identify the problematic content as soon as possible to prevent it from going viral.

4.2.4.2. Content that requires human review, because it violates some but not all of the common criteria, can be forwarded to human moderators.

4.2.5. OSPs should provide regular training on current societal developments to their human content moderators, and if possible also to the engineers working on content-related projects. Currently little is known about how OSPs like Facebook train their human content moderators.

4.2.5.1. According to Kate Klonick, human content moderators receive personal training to ensure that they enforce harmonized rules and not their own cultural values and norms.¹⁵⁰

4.2.5.2. According to leaked documents, published mainly by online media, the material taught in these courses is modified to keep up with current events, such as after Charlottesville.¹⁵¹

4.2.5.3. At the same time, according to a recent lawsuit against Facebook, content moderators, despite their training, are prone to trauma after reviewing thousands of videos, images, and live-streamed broadcasts of child abuse, rape, torture, bestiality, beheading, suicide, and murder.¹⁵² Some content moderators have committed

150 Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

151 Joseph Cox, *Leaked Documents Show Facebook's Post-Charlottesville Reckoning with American Nazis*, MOTHERBOARD, May 25, 2018; Angwin & Grassegger, *supra* note 99.

152 *Facebook Failing to Protect Moderators from Mental Trauma, Lawsuit Claims*, THE GUARDIAN, September 25, 2018.

suicide.¹⁵³ For example, Google limits its YouTube content moderators to four hours of disturbing content a day.¹⁵⁴

4.2.6. OSPs should adjust the composition of their content-moderation staff to reduce bias and ensure diversity. A mix of trained personnel from different cultures and languages can improve the content moderation department's ability to implement the common criteria for identifying hate speech in a given context.

4.2.7. As mentioned above, algorithmic decision-making remains limited and imperfect. Hence we recommend that the automated process only flag content for human decision-making and not remove content without human intervention. This provision can and should be reexamined as machine-learning technologies advance.

4.3. Step 2: Notification of Violations

4.3.1. OSPs should make it possible for law-enforcement agencies to notify them of violations of the criteria. Some OSPs, such as Facebook and Twitter, have published guidelines on how law-enforcement agencies can notify them about problematic content and instituted dedicated mechanisms to request information and to submit takedown requests.¹⁵⁵ This mechanism may require law-enforcement agents to identify themselves before they can obtain access.¹⁵⁶ OSPs have recently begun publishing transparency

153 *The Cleaners* (Gebrueder Beetz Filmproduktion) (2018).

154 Nick Statt, *YouTube Limits Moderators to Viewing Four Hours of Disturbing Content per Day*, THE VERGE, March 13, 2018.

155 See Twitter, *Guidelines for Law Enforcement*; Facebook, *Information for Law Enforcement Authorities*; Google, *Transparency Process for User Data Requests FAQs*.

156 See, e.g., Twitter, *Legal Request Submissions: Please Confirm your Identity*; Facebook, *Law Enforcement Online Requests*; Uber, *Law Enforcement Portal Overview*.

reports about the requests received from law-enforcement agencies.¹⁵⁷ Given the existence of these co-regulatory mechanisms, we suggest maintaining and possibly updating these channels of communication. These notifications should be channeled through national contact points designed jointly by OSPs and law-enforcement agencies and be given priority treatment, as defined in Step 4.

4.3.2. Civil society organizations and OSPs should strengthen their partnerships, provide each other with information about flagging mechanisms and organizational policies, and work to extend the geographical spread of their partnerships. OSPs should permit more civil society organizations to act as “trusted reporters” who flag content that allegedly violates the common criteria. YouTube has a “Trusted Flagger” program in which it provides robust mechanisms for notifying it of content that violates its Community Guidelines. These mechanisms include a bulk-flagging tool for multiple simultaneous reports, private forum support, visibility of decisions on flagged content, and prioritized reviews.¹⁵⁸ Currently, dozens of civil society organizations are acting as trusted reporters.¹⁵⁹

4.3.3. Creating a user interface for submitting complaints:

157 See [Google's transparency report](#); [Facebook's transparency report](#); and [Twitter's transparency report](#).

158 According to YouTube, to be eligible flaggers must flag frequently, have a high rate of accuracy, and attend a training course on YouTube's guidelines and enforcement processes. See YouTube, [YouTube Trusted Flagger program](#).

159 The report of the European Commission lists 33 civil society organizations that act as trusted reporters. There was only a 65.6% removal rate for notifications using trusted flaggers/reporters channels. See European Commission, [Code of Conduct on Countering Illegal Hate Speech Online: One Year After](#) (June 2017). For more information see YouTube's Trusted Flaggers Program.

4.3.3.1. OSPs should provide users with a flagging mechanism incorporated into the standard user interface.

4.3.3.2. Although it can be a bit cumbersome, providing granular information on a case reported is a requirement that helps the OSP reach a decision about the case more quickly, and on the basis of relevant information. It also makes it easier to distinguish true from false claims. Google recommends that users provide more details to help it identify the content, add voiceover or text narration to explain it, and state what users should not do with online content.¹⁶⁰

4.3.3.3. We recommend that OSPs require notifiers to assist them, as much as possible, in dealing with the factors involved in the company's implementation of the common criteria.

4.3.3.4. The notification mechanism should ensure that the OSP is made aware of the complaint immediately. In any case, the initial acknowledgment that the complaint was received should be sent within 24 hours.¹⁶¹

4.3.3.5. While many OSPs provide a complaints mechanism,¹⁶² too many locate it in a hard-to-find location at the bottom of pages or hidden behind several web-clicks, or require filling in a form and copying over the address of the original

160 YouTube, *Guidelines for Adding Content*.

161 According to the European Union, in 51.4% of the cases, OSPs assessed notifications in less than 24 hours, in 20.7% in less than 48 hours, and in 14.7% in less than a week. In 13.2% of the cases it took the OSP more than a week to assess a notification. See European Commission, *Code of Conduct on Countering Illegal Hate Speech Online: One Year After* (June 2017).

162 See Appendix B for examples of the types of flagging mechanisms offered by OSPs.

post. Frequently users have to submit an email, which makes filing a complaint much more difficult.

4.3.3.6. We recommend that flagging mechanisms be integrated into the main user interface, directly accessible, and in a standard location with an easily recognizable button.¹⁶³

4.3.3.7. The mechanism should not be accessible only from a different webpage and should not require leaving the area of the questionable content.

4.4. Step 3: Organizational Decision

4.4.1. After receiving a removal request and before deciding about the relevant content, the OSP should contain the content to limit its virality. Different platforms implement this function in different ways:

4.4.1.1. YouTube has rules about which content can earn revenue for creators and has launched new comment-moderation tools (including shutting comments down altogether).¹⁶⁴

4.4.1.2. YouTube, Twitter, and Facebook have all started using mechanisms that warn users or block access to offensive and extreme videos and pictures. Users who want to access these videos or pictures must click on the picture or on a button next to it to access it, thus affirming their informed consent to exposure to the offensive material.

¹⁶³ Similar demands are found in Section 3(1) of the German Network Enforcement Act.

¹⁶⁴ Wojcicki, *supra* note 109.

4.4.1.3. Although this policy for comment-moderation tools and user consent is appropriate and should continue, it shifts responsibility to users. Our recommendation, on the other hand, is that OSPs draft a policy that bears directly on the content-distribution algorithms. As compared to removal of content, this algorithm-based process is less injurious to users' freedom of expression and can also be used as an intermediate solution until a final decision is made.

4.4.2. The common criteria can help the OSP identify hate speech and decide on differential responses to content, based on its severity.

4.4.2.1. Using the common criteria, the OSP can develop algorithm-based or human-based responses as a function of the content's severity and the extent to which it violates the common criteria implemented by the company.

4.4.2.2. A company can decide that content that violates the strictest definitions will be automatically deemed to be "manifestly unlawful content," automatically flagged for human reviewers, and removed. Content that is less severe should be flagged for human reviews or require users' consent to watch it.

4.4.2.3. Content that the OSP identifies as falling on the more lenient sides of the different criteria can require additional human intervention and consideration by the different corporate tiers.

4.4.3. OSPs should decide on the extent of the restriction as a function of the origin of the request.

4.4.3.1. Requests made by the national authorities or law-enforcement agencies: On the one hand, as state actors, law-enforcement agencies are expected to consider content in a broader context that is subject to democratic safeguards, balancing the various public interests involved

against a takedown request, including public order and safety, freedom of expression, and other civil liberties. On the other hand, there are public and democratic concerns that content-removal requests may target content that the government dislikes.

4.4.3.1.1. OSPs should consider these two perspectives and develop a response model for each country.

4.4.3.1.2. Based on its policies for a particular country and its experience with its law-enforcement agencies, OSPs can select the severity of the content restriction applied. They can remove the content, limit its virality, or ask for a court order to remove it.

4.4.3.1.3. The OSP can decide to limit the content's virality on a national level (geo-block) instead of on a regional or global scale.

4.4.3.1.4. An OSP may decide that law-enforcement agencies need to train their personnel with the company before establishing reliable notification channels.¹⁶⁵

4.4.3.1.5. For further details on possible responses, see §4.4.5.

4.4.3.2. Requests made by trusted reporters affiliated with civil society organizations: On the one hand, in many cases OSPs may decide that specific civil society or non-governmental actors are worthy of becoming trusted reporters.¹⁶⁶ On the other hand, with trusted reporters, unlike law-enforcement agencies, there is no external oversight or possibility of requesting a court order.

165 For further information see [YouTube's Trusted Flaggers program](#).

166 *See id.*

4.4.3.2.1. This means that the flagging of content by trusted reporters can lead to a decision to block content but require some form of secondary confirmation by algorithmic or human moderation.

4.4.3.2.2. Whereas content flagged by law-enforcement agencies can be geo-blocked for a specific country, a flag by a trusted reporter can help the OSP decide whether to limit the virality of content on a regional or global scale.

4.4.3.2.3. OSPs should also train civil society organizations in fulfilling their “trusted reporter” role. This training can help the company get to know the organization and determine whether a more specific policy should be associated with complaints coming from a particular civil society organization.

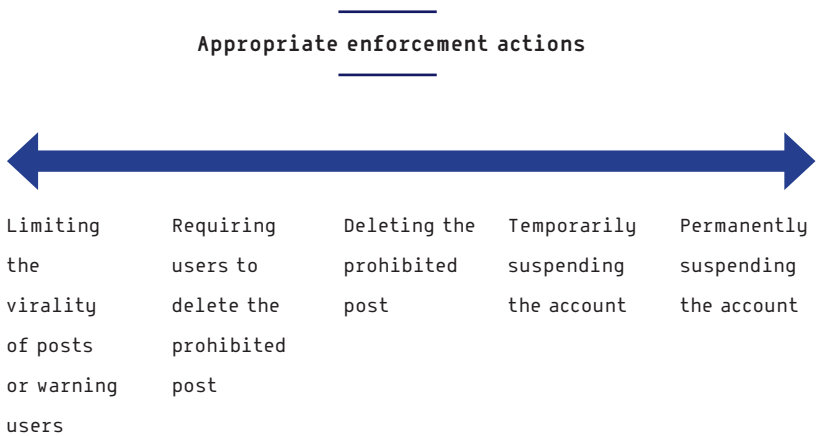
4.4.3.3. Requests from users: Like trusted reporters from civil society organizations and requests by law-enforcement agencies, users, too, may report content they find harmful or inappropriate to the social network. Because of the greater likelihood of false claims or the dependence on other factual circumstances, OSPs should develop a policy that limits the recourse to algorithmic decision-making. Instead, their policy should include more human-based content moderation and lead to less severe responses than to requests filed by law-enforcement agencies and civil society organizations. For instance, although the German Telemedia Act mentions the possibility of contacting the user who posted the content,¹⁶⁷ Twitter, because it accepts reports from anyone, states that

167 See §3(2).3 of the German Network Enforcement Act.

it needs to hear directly from the target to ensure it has the proper context.¹⁶⁸

4.4.4. In the wake of a decision by the OSP that the content does in fact violate its policies, it should choose among several enforcement actions. These range from steps to limit the post's virality (for instance, to limit the virality of content that spreads misinformation or can dehumanize or legitimize hostile actions over time), to the removal of the content from the entire platform, and finally permanent suspension of the user's account.

4.4.5. The severity of possible responses is described by the next scale:



4.4.5.1. OSPs employ algorithms to *limit the virality of questionable posts*. Another option is to warn users that the content may be disturbing and require their

consent to watching it. Both Facebook and Google use this mechanism.¹⁶⁹

4.4.5.2. In addition to limiting the virality of posts, OSPs can warn users that their content violates their TOS or community guidelines and *require the users to remove the content* themselves by a stated deadline. Twitter specifies that users may be required to remove an offending tweet before they are allowed to tweet again.¹⁷⁰

4.4.5.3. Going beyond the previous option, the OSP can *delete the content itself* instead of leaving the decision to the user who posted or reshared it. Several platforms have policies that allow them to remove content without waiting for the user to act.¹⁷¹

4.4.5.4. An OSP can decide to *temporarily suspend the account* of a user who infringes its policies. This sanction is especially relevant for users who have repeatedly violated the policies or have not responded to the OSP's direct communication regarding their actions. According to Twitter, it may temporarily suspend accounts until a user deletes offending tweets.¹⁷²

4.4.5.5. OSPs can decide to *permanently suspend a user's account*. This sanction is especially relevant for users who have posted manifestly unlawful content several times and after all other actions have failed to get them to change their online practices.

169 Allan, *supra* note 140; Wojcicki, *supra* note 109. Facebook has a similar policy for graphic violence.

170 Twitter, [Hateful Conduct Policy](#).

171 See [Facebook's hate-speech policy](#).

172 Twitter, [Hateful Conduct Policy](#), *supra* note 170.

For instance, after the removal of Alex Jones’s Info-Wars page in August 2018, Facebook explained its account suspension policy.¹⁷³ According to Facebook, every time Facebook removes content that violates its community standards, it chalks up a demerit against the user, and, if it was on a page, for that page as well. Facebook will suspend users based on the severity of the violation. First-time offenders receive a warning. If they continue, Facebook may temporarily block their account, thus restricting their ability to post. Extreme content and repeat offenders will be suspended immediately. For pages, after a certain threshold, which Facebook does not specify, it will “unpublish” the entire page. Pages can appeal the decision to unpublish them. If the page owners do not appeal or their appeal is rejected by Facebook, the page is permanently removed.

4.4.6. Additional steps, not included in the scale, can address the user being attacked or targeted. These steps include informing the user, offering assistance, providing information on where users can receive information or support (mainly from members of trusted reporter lists), or contacting law-enforcement agencies. These steps should apply especially when law-enforcement agencies did not initiate the report.

4.4.7. Timetables and notification of action:

4.4.7.1. A decision about manifestly unlawful content should be made within 24 hours, unless the law-enforcement agency agrees to a longer timeframe.

173 *Enforcing Our Community Standards*, FACEBOOK NEWSROOM, August 6, 2018.

4.4.7.2. A decision about blocking or removing unlawful content should be made within seven days of the submission of the complaint.

4.4.7.3. A longer delay may be allowed if the decision regarding the content depends on whether a factual allegation is false or on other factual circumstances. In such cases, the OSP can give the user an opportunity to respond before reaching a decision; in the case of a request by a law-enforcement agency it can ask for a court order.¹⁷⁴

4.4.8. After the decision, the law-enforcement agency or person who filed the notification about the content should be informed of the decision—individuals through their user accounts and law-enforcement agencies through the national contact points. The OSP should keep a record of the content involved, of its decision, and of the measures taken (including removal).¹⁷⁵

4.4.9. Based on the severity of the content and the company's decision, the OSP can provide users whose content was blocked or removed with information about the decision.¹⁷⁶ Notification of a decision to remove content or suspend an account should include at least the following details:¹⁷⁷

4.4.9.1. Sufficient information to identify the content concerned.

174 A similar mechanism exists in §3 of the German Network Enforcement Act.

175 The requirement is within the scope of Directive 2000/31/EC.

176 For YouTube's appeal procedure, see [Appeal Community Guidelines actions](#).

177 Based on *The Santa Clara Principles on Transparency and Accountability in Content Moderation*.

4.4.9.2. The specific clause in the company's policies that the user violated.

4.4.9.3. If possible, and unless prohibited by law, how the content was detected and removed. The identity of individual flaggers and civil society organizations should not be revealed. Law-enforcement agencies can be identified, unless this is prohibited by law.

4.4.9.4. Whether the user can appeal the decision.

4.4.9.5. An appeal mechanism provided as part of a set of transparent policies and mechanisms. At minimum, the appeal process should include the following:¹⁷⁸

4.4.9.5.1. A human reviewer or a panel of reviewers that was not involved in the initial decision. The use of independent external reviewers should be deemed a component of the content removal process.

4.4.9.5.2. An opportunity for the user to submit additional information for consideration in the review.

4.4.9.5.3. The option to modify the content and add context in a way that permits its publication

4.4.9.5.4. Notice of the outcome of the review and a statement of the reasoning sufficient to allow the user to understand the decision.

4.4.10. Additional accountability and transparency mechanisms for the OSP's decision are presented below in Step 4.

4.5. Step 4: Transparency and Accountability Mechanisms

4.5.1. OSPs should ensure that a thorough explanation of how they implement the material hate-speech criteria is available to users in the platform TOS and community standards document. The exact internal procedures for implementation of the hate-speech criteria can remain confidential so as to prevent their being gamed.

4.5.2. Hate-speech complaints should be monitored on a monthly basis.

4.5.2.1. This requirement can be filled by a member of the OSP's senior management or by personnel specifically assigned to do so, provided they have a direct line of communication to senior management. If no one has been tasked with this responsibility, it falls to either the CEO or the General Counsel to address the relevant policies.

4.5.2.2. Though there are calls to create an external oversight or appeal mechanism for content moderation, we consider this mechanism to be highly dependent on the OSP's economic capacity and platform size. What might work for Facebook might not work for smaller platforms. For the latter, monthly managerial oversight and transparency reports can suffice.

4.5.3. The internal monitoring of complaints should include all requests made. The OSP should analyze the requests according to their location on the common-criteria scales, origin, number, the time it took to process them, and the final decision taken.

4.5.4. Collecting data on posts: For every content item marked as infringing the OSP's hate-speech policy, it should collect data on the shareability of that content at that time. The data should include how many likes or views the content received and how many

times it was shared or re-tweeted. If the content was flagged but not removed, the OSP can also collect data on the content going forward.

Specific consideration should be given to the following cases and should be mentioned in the transparency reports:

4.5.4.1. Flagged content was not found to violate the OSP's policies, but the content moderation team decided to remove it from the platform.

4.5.4.2. Flagged content was found to violate the OSP's policies, but the content moderation team decided not to remove it from the platform.

4.5.5. To assist in the training of future staff and help senior management with policy development, the report should include case studies. These should note the relevance of the common criteria as implemented by the OSP as well as how the company made its final decision. The case studies should also refer to instances in which the moderators found it difficult to decide whether hate speech was involved or how to apply the corporate policies. If the OSP noted any deficiencies in handling the case, relevant senior management should be notified and find ways to rectify them.

4.5.6. OSPs should provide information in the form of transparency reports, based on the information described below, and specifically on the handling of complaints about unlawful content. The reports should be easily recognizable, directly accessible, and permanently available, for instance by posting to a designated webpage. The reports should include at least the following:

4.5.6.1. A summary of the OSP's efforts to eliminate hate speech from its platform: The summary should include a broad description of the company's policies on

implementing the material criteria as well as the statistics found in the report to the management.

4.5.6.2. A description of the mechanisms for submitting complaints and the criteria applied when deciding whether to delete or block unlawful content.

4.5.6.3. The number of incoming complaints, broken down by who submitted them and the reasons for the complaint.

4.5.6.4. The number of complaints in the reporting period that resulted in the deletion or blocking of content, and either permanent or temporary suspension of users for violations of content guidelines. These data should be broken down as follows:¹⁷⁹

4.5.6.4.1. The total number of discrete posts and accounts that were flagged.

4.5.6.4.2. The total number of discrete posts that were removed and of accounts that were suspended.

4.5.6.4.3. How many discrete posts and accounts were flagged, and how many discrete posts were removed and accounts suspended, by category of rule violated.

4.5.6.4.4. How many discrete posts and accounts were flagged, how many discrete posts were removed, and how many accounts were suspended, by content format.¹⁸⁰

4.5.6.4.5. How many discrete posts and accounts were flagged, how many discrete posts were

179 *Id.*

180 *E.g.*, text, audio, image, video, live stream.

removed, and how many accounts were suspended, broken down by the source of the flag.¹⁸¹

4.5.6.4.6. How many discrete posts and accounts were flagged, how many posts were removed, and how many accounts were suspended, broken down by the location of the flaggers and the users affected.

4.5.6.4.7. How long it took to take down content that was the subject of complaints.

4.5.6.5. Information about notifications and the disabling of access to or removal of illegal online hate speech.¹⁸²

4.5.6.6. The measures employed to inform the relevant bodies or persons of the decision made.

4.5.6.7. Information about training and support of the persons responsible for processing complaints.

4.5.6.8. To enable future research, the data reported should be provided in a regular (ideally quarterly) report, in an open-license machine-readable format.¹⁸³

4.5.7. In addition to the training programs for content moderators, the OSP's management should make sure that the moderators have access to counseling and support programs.¹⁸⁴

181 *E.g.*, governments, trusted flaggers, users, different types of automated detection.

182 Such reports would enable law-enforcement agencies and civil society organizations to familiarize themselves with the methods for identifying and notifying OSPs of violations of the common criteria.

183 *See Santa Clara Principles*, *supra* note 177.

184 Similar support programs are required under §3(4) of the German Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act).

4.5.8. OSPs should provide information about their collaborations with civil society organizations recognized as trusted reporters, as well as about how users can contact these organizations.

4.5.9. OSPs should cooperate among themselves to enhance and share best practices.¹⁸⁵ This collaboration can lead to a code of conduct, a shared closed list of unaccepted terms or symbols, external certification schemes or dispute-resolution bodies, or technological solutions. All such cooperation should include, to the extent possible, the views of supranational actors such as the European Commission and of civil society actors.

4.5.10. Based on the internal and external reports, each company's senior management should assess and update its material and procedural implementation of the co-regulatory mechanism on a regular basis. The OSP should also review its transparency mechanism.

4.5.11. OSPs should use their platforms to educate users and raise their awareness about the types of content that are not permitted under their rules and community guidelines. Attention should be paid to ways of reaching users who are not familiar with the notification system. One possibility is to run joint educational programs with civil society organizations or state actors.

185 In October 2017, it was reported that the Anti-Defamation League had joined Facebook, Twitter, Google and Microsoft, among others, to curb online hate speech. As part of a Cyberhate Problem Solving Lab, OSPs will exchange ideas and develop strategies to try to curb hate speech and abuse. See Peter Strain, *Anti-Defamation League, Tech Firms Team to Fight Online Hate*, cnet.com, October 10, 2017. See also IP/16/1937, European Commission, *supra* note 3. See Code of Conduct, *supra* note 3.

Chapter 5

Advantages and Disadvantages of the Proposed Co-Regulatory Model

The proposed common criteria and procedures should be adopted and implemented by OSPs as part of their broader corporate governance scheme and, more specifically, their content-moderation policy. They can do this in various ways. One option is for supranational or national legislation to mandate the implementation of content moderation. Another possibility is a self-regulatory mechanism. Co-regulation is the third option.¹⁸⁶ In the following paragraphs we discuss the advantages and disadvantages of each model in order to highlight why we consider the co-regulatory model to be the best of the three.

The main advantage of national legislation for the regulation of hate speech is that it can achieve a balance among local normative, constitutional, moral, and social values, such as the right of free expression and public order and safety. This balance can come through legislation that assigns OSPs direct responsibility for content posted on their platforms, legislation that requires them to moderate content, or court orders or warrants that require them to delete content. Governments would block access to the products and services of OSPs that do not comply or fine them. Accordingly, each country could debate the appropriate balance between individual freedom of expression, off- and online, and other social values, reflecting their own unique conditions and population.

At the same time, legislation carries several disadvantages. National legislation is not really able to deal with the global character of the internet. National laws that do not coincide with international or supranational

186 While chapter 3 also discusses information-based mechanisms, we utilize them in the context of chapter 4(c) to support the co-regulatory model with transparency and accountability mechanisms rather than as a stand-alone model.

conventions create online islands of national jurisdiction. These islands change the nature of the internet as a global medium of communication and create tension and sometimes contradictions between different jurisdictions. In addition, whereas the internet and its information and telecommunication technologies develop rapidly, legislation can take a long time to find the right balance and then be enacted. This gap between the law book and current technology may be hard to close, even if authority is delegated to law-enforcement agencies and the courts. As a result of these disadvantages, OSPs may decide to geo-block specific services from a country or decide that it is simpler to apply the stricter rules to countries with more lenient legislation. For hate speech, when directed against minorities, geo-blocking and strict implementation of global rules can lead to the use of VPNs to bypass the geo-blocking and reach otherwise inaccessible content. The result could be a race to the bottom on both the global and the national levels.

Self-regulation by OSPs has several advantages. The most important is that because OSPs are multinational corporations, their self-regulation has transnational effect. Corporate decisions, and especially the technologies developed as a result, can reach every country where a company provides services. Similarly, when the OSP implements self-regulation, it can harmonize the rules across all the countries it serves. Lastly, it is more difficult to circumvent self-regulatory than national legislation. If an OSP takes down content, it is easier for it to do so automatically across the platform. Users who want to access the content or who have been kicked off the platform must find another platform.

But self-regulation also has disadvantages. OSPs and their self-regulatory practices do not enjoy the normative legitimacy needed to balance values. This is especially true given the economic interests involved, which limit OSPs' desire to regulate themselves in a way that can balance the different markets they serve. For instance, if OSPs do not regulate hate

speech, users may leave the platform; if users leave, advertisers and app developers will soon follow.

To summarize the foregoing: On the one hand, national legislation keeps the normative decision with government and state actors and away from private OSPs. OSPs have an incentive to comply with the law. On the other hand, there is a clear benefit to rules adopted by multinational corporations and implemented across national borders; they can adapt to new technologies more quickly and have transnational implementation with a harmonizing effect.

The third model, which we presented, is co-regulation. Co-regulation carries with it many of the advantages of the first two models, because public and private actors share responsibility and work together to achieve public goals. At the same time, while law-enforcement agencies are national, co-regulation does not have to be: OSPs can still implement co-regulation globally. Two disadvantages have to be mentioned. First, co-regulation does not always work, especially when the private sector has no incentive to implement it. Second, in order to achieve necessary compromises, co-regulatory schemes can be ambiguous. This ambiguity may leave ample room for interpretation by the OSPs that keep them within the scheme, but it can also lead to difficulties in creating clear and agreed-upon rules, practices, and implementation.

Our co-regulatory model has several advantages. First, it takes the normative principles for regulating hate speech that are standard in comparative law and makes them the common criteria. Specifically, we chose to adopt these criteria because we believe that a more specific definition is required, one that is based on national criminal legislation, global conventions, regional agreements, and OSPs' policies. As such, our model maintains the normative and moral balance regarding hate speech that exists in most Western countries. This path makes it possible for us to identify the common mechanism and avoid the consequences

of national laws that do not correspond to the practices of global social-media providers.

Second, our model builds on the fact that platforms moderate content,¹⁸⁷ and in so doing decide what the regulatory rules are. We believe, however, that there are sufficient public and private interests to change the course of hate speech online, and on social networks' platforms in particular. This is why our model provides a general benchmark using a co-regulatory model—one that includes OSPs, law-enforcement authorities, and civil society. We do so without challenging constitutionally protected rights or suggesting that existing legislation be amended. On the other hand, if OSPs lack the incentive to act, governments can use the legal and quasi-legal mechanisms we mentioned in chapter 3.

Third, our model includes procedures for implementation of the common criteria by OSPs. We believe that a model based on scales can help companies implement the policies through human moderators, technology-based content monitoring, and algorithmic flaggers. Additionally, the scales model permits OSPs and their management to determine whether their policies are too lenient or too strict and move along the scales in search of a different policy. Our model does not make any assumptions about an OSP's corporate size, technological capabilities, or deep pockets. Our model can be used by both OSPs of different sizes—from huge to small and medium-sized enterprises (SMEs)—while leaving it to the OSPs to determine their position on the criteria and how they need to address the procedural aspects of the proposed model.

Fourth, our model offers a shared terminology based on the common criteria and implementation procedures, and includes accountability and transparency mechanisms relating to the enforcement policy

187 TARLETON GILLESPIE, *CUSTODIANS OF THE INTERNET* (2018).

implemented by the OSPs. While our model is open to the criticism that it can lead to censorship or to over-lenient policies that governments and other political actors believe approve too much content, this debate about lenient or strict policies for content moderation can move forward only if law-enforcement agencies and civil society actors can compare the different platforms, especially with regard to how they implement the common criteria and how strict or lenient their policies are.

Our model does have its disadvantages. For the most part, it relies on the belief that both governments and OSPs are motivated to implement it. In addition, the model could be cumbersome (in comparison to current self-regulatory policies), because it includes sub-definitions and scales. Furthermore, it is based on knowledge of current OSP policies and information from leaked documents. As such, it might be insufficiently dynamic for self-regulation (though preferable to legislation) and require updating as new technological and algorithmic capabilities are developed. Lastly, we are aware that some content-moderation issues, such as the liability of the administrators of forums, closed groups, and pages, remain outside the model.

However, we consider our co-regulatory mechanism to be the best available one in the current circumstances. Applying a shared jurisdiction with common criteria can lead to harmonization and help countries and users understand the extent to which each platform follows the norms for regulating hate speech. If all—and most importantly the largest—platforms implement the model, each platform could display its policy choices; regulators and users could use this policy to decide which platform to use and how to respond when national regulation is required. On the other hand, self-regulation mechanisms lack democratic legitimacy, do not involve law-enforcement agencies, and limit platforms' ability to collaborate where needed. A co-regulation mechanism can overcome these limitations. In our view, the model is easy to implement, makes possible international agreement about the required balance

while maintaining corporate flexibility, and enables users to choose by providing them with knowledge that empowers them.

Our model incorporates decision-making by humans or algorithms. The decision to incorporate human or algorithmic decision-making may vary from company to company and from department to department. Appendix B offers examples of how several major OSPs practice content moderation. These companies can afford to develop algorithm-based content moderation or to hire human moderators on a scale that might not be possible for smaller companies with limited resources. We do not expect all companies to implement the same mechanisms and the same method. However, the implementation steps can help executives understand what measures they should think about when they develop procedures for content moderation.

Algorithmic decision-making has many advantages and disadvantages. On the one hand, artificial intelligence for content moderation can resolve crises on a global scale, while helping OSPs like Facebook deal with questions of censorship, fairness, and moderation by humans. The primary benefit of algorithmic decision-making is the speed of the decision about massive quantities of content. According to Mark Zuckerberg, artificial intelligence can solve content-moderation problems such as hate speech, terrorist propaganda, and fake news.¹⁸⁸ In April 2018, however, Zuckerberg asserted that it would take Facebook five to ten years to develop artificial intelligence for content moderation with enough accuracy to flag potential risks.¹⁸⁹ For now, companies such as Google and Facebook are known to

188 Drew Harwell, *AI Will Solve Facebook's Most Vexing Problems, Mark Zuckerberg Says. Just Don't Ask When or How*, WASHINGTON POST, April 11, 2018.

189 *Id.* Meanwhile, algorithms are used to flag content. For instance, According to Google's Transparency Report, 74.2% of content removed

use algorithms only to flag content for referral to human decision-making; the algorithms do not remove content without human intervention.

On the other hand, scholars claim that artificial intelligence is a “MacGuffin” designed to solve Zuckerberg’s and other executives’ liability problem.¹⁹⁰ In fact, the technologies’ state of maturity, accuracy, and scalability are all factors that might affect a future decision to rely on algorithmic and specifically NLP technologies to identify hate speech. In addition, algorithmic decision-making challenges democratic rights. The delegation of responsibility to algorithms means less accountability and less transparency and makes it more difficult to ferret out discrimination caused by hidden manipulations.¹⁹¹ In a nutshell, algorithms have biases and may not be able to include all relevant cultural and legal aspects and context in their decision. Although companies themselves are not always transparent about their policies, algorithms take opaque decision-making a step further, because users and coders may not understand the reason behind a decision. Scholars also worry that even transparent algorithms may produce discriminatory results, and thus offer transparency of inputs and open-sourced code.¹⁹²

from YouTube was first flagged through the automated flagging mechanism. See [Google's Transparency Report](#).

190 See James Grimmelmann's response at the [washingtonpost.com](#) website.

191 FRANK PASQUALE, *THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION* (2015).

192 Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017).

Appendix A

Defining the Hate-Speech Policy Problem

Although it is easier today to characterize the consequences of hate crime and xenophobia,¹⁹³ state institutions still find it difficult to define and identify them.¹⁹⁴ Because of the lack of a definition accepted by different countries and platforms, and of standard record-keeping procedures, among other things, policymakers have insufficient information and are unable to fully comprehend the scale of the phenomenon. Furthermore, the absence of precise information poses a challenge to the development of data-driven policies to combat hate crime and xenophobia and makes it difficult to assess the policies' effectiveness. The lack of reporting prevents the police and courts from investigating and prosecuting hate crimes and complicates the ability of welfare and medical systems to assist victims.

Despite its importance for policymaking and for the justice and welfare systems, the collection of data about hate crime and xenophobia has been limited; often what is available cannot be compared and consolidated, because of different collection and classification methodologies.¹⁹⁵ The

193 Hate crimes harm people's physical and mental health as well as violate their fundamental rights, including the rights to human dignity, equality of treatment, and freedom of thought, conscience, and religion.

194 For instance, the FRA data show that only a few EU member states record antisemitic incidents in a way that allows them to collect adequate official data. This failure to record hate crimes, coupled with victims' hesitance to report incidents, leads to gross underreporting of the extent, nature, and characteristics of antisemitic and other hate crime in Europe. See FRA, DISCRIMINATION AND HATE CRIME AGAINST JEWS IN EU MEMBER STATES: EXPERIENCES AND PERCEPTIONS OF ANTISEMITISM (2013).

195 There are different methodologies among European countries. This has spurred the FRA to convene a subgroup of experts and professionals within the European Union High Level Group on Combating Racism, Xenophobia and Other Forms of Intolerance. This group helps member states develop a common methodology for data collection and recording of hate crimes. See *id.* at 6.

next few paragraphs present data about online hate crime and xenophobia and about the methods used to collect the data.

Although several national and supranational agencies collect official data from local police and court records, these data cannot always be compared. In Europe, for instance, the data published by the European Union Agency for Fundamental Rights (FRA) indicates that antisemitism—a form of hate speech that is particularly sensitive in the European context—is a matter of grave concern there;¹⁹⁶ but there are gaps in the data and under-reporting.¹⁹⁷ For instance, the FRA notes that the OSCE Office for Democratic Institutions and Human Rights (ODIHR) collects data from all 28 EU member states for input to an online crime-report database. The data collected from governmental sources, civil society, and intergovernmental organizations relates to “bias motivations,” one of which is antisemitism.¹⁹⁸ So although the FRA can present data on each of the European member states,¹⁹⁹ the data collected by the European

196 For instance, given the lack of a standardized methodology, sometimes even within a single state over time, “it cannot be assumed that antisemitism is necessarily more of a problem in Member States where the highest numbers of incidents are recorded than in those where relatively few incidents are recorded” (*id.*, at 85).

197 According to the FRA, “evidence collected by FRA consistently shows that few EU Member States record antisemitic incidents in a way that allows them to collect adequate official data.” Also, the data that do exist “are generally not comparable, not least because they are collected using different methodologies and from different sources across EU Member States” (*id.* at 5).

198 European Union Agency for Fundamental Rights, [Antisemitism, Overview of Data Available in the European Union 2006–2016](#) (November 2016). See also the European Commission against Racism and Intolerance (ECRI), [Annual Report on ECRI's Activities: Covering the Period from 1 January to 31 December 2016](#), CRI(2017)35 (June 2017).

199 For instance, the official data of EU member states show that, in 2015, the United Kingdom, France, the Netherlands, and Germany had 786, 715, 428, and 192 antisemitic events, respectively (*id.*).

institutions cannot be compared due to gross under-reporting of the extent, nature, and characteristics of antisemitic incidents in Europe. As a result, the FRA can provide only an overview and its data cannot be taken as an accurate portrayal of the prevalence of antisemitism in any particular EU member state.

In the United States, the Federal Bureau of Investigation (FBI) collects data on hate crimes through the Uniform Crime Reporting (UCR) program.²⁰⁰ The data for 2015 indicate that 59.2% of the 5,818 single-bias incidents, with 7,121 victims, were motivated by racial, ethnic or ancestry bias; 19.7% were prompted by religious bias.²⁰¹ On both sides of the Atlantic, “official” data are collected from official authorities, but the collection, recording, and display processes suffer from gaps, inaccurate classification, and a lack of standardized categorization. To supplement data on the activities of law-enforcement agencies, several methodologies have been developed to define, present, and display changes in online hate crime over time.

One such policy was introduced after the adoption of the Code of Conduct on Countering Illegal Hate Speech Online²⁰² by the European

²⁰⁰ According to the FBI, 14,997 law-enforcement agencies participated in the Hate Crime Statistics Program in 2015. Of them, 1,742 agencies reported 5,850 hate-crime incidents involving 6,885 offenses. See U.S. Department of Justice, Federal Bureau of Investigation, *Uniform Crime Report, Hate Crime Statistics, 2015* (released Fall 2016).

²⁰¹ Additional data showed that of 6,837 single-bias hate-crime-related offenses, 58.9% were motivated by racial, ethnic, or ancestry bias, and 19.8% by religious bias. Also, out of the 4,029 race-motivated hate crimes, 52.7% were directed against African Americans; 51.3% of the 1,354 hate crimes reported based on religion were directed against Jews and 22.2% against Muslims. See *id.*

²⁰² IP/16/1937, European Commission, *supra* note 3; Code of Conduct, *supra* note 3.

Commission, Facebook, Twitter, YouTube, and Microsoft, as well as the implementation of Framework Decision 2008/913/JHA regarding online contexts.²⁰³ In the second evaluation exercise, conducted between March and May 2017, 31 organizations and three public bodies reported on a sample of 2,575 notifications submitted as part of the Code of Conduct.²⁰⁴ The EU noted significant progress by social-media platforms, mainly that social networks have become more efficient and faster in assessing notifications.²⁰⁵ The platforms have also strengthened their systems for reporting illegal hate speech and trained their staff.²⁰⁶ According to the European Commission's Directorate-General for Justice and Consumers, "cooperation between IT companies and civil society organizations leads to a higher quality of notifications, more effective handling times, and better reactions to notifications."²⁰⁷ Nevertheless, the EC believes that there is still room for improvement in the platforms' transparency and

203 EU Council Framework Decision 2008/913/JHA (3) on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law.

204 European Commission, Code of Conduct on Countering Illegal Hate Speech Online: One Year After (June 2017).

205 *Id.*

206 According to the EC's findings, "Overall, 1522 of the notifications (59.1%) led to the removal of the notified content, while in 1053 cases (40.9%) the content remained online. Facebook removed the content in 66.5% of cases, Twitter in 37.4% and YouTube in 66% of the cases. This represents a substantial improvement for all three companies compared to the results presented in December 2016, where the overall rate was 28.2%" (*id.* at 2).

207 According to the EC's findings, "In 51.4% of cases IT companies assessed notifications in less than 24 hours, in 20.7% in less than 48 hours, in 14.7% in less than a week and in 13.2% it took more than a week. Facebook assessed the notifications in less than 24 hours in 57.9% of the cases and in less than 48 hours in 24.9% of cases. The corresponding figures for YouTube are 42.6% and 14.3% and for Twitter 39% and 13.7%, respectively. There is a positive overall trend in the

feedback systems.²⁰⁸ In January 2018, it published the results of its third evaluation, carried out in November and December 2017. This revealed further progress: IT companies removed 70% of the illegal hate speech brought to their attention and reviewed an average of 81% of such notifications within 24 hours.²⁰⁹

While the public sector focuses on the broad identification of hate speech, private organizations and institutions that try to analyze and quantify hate speech concentrate on attacks that target a specific group or groups. For instance, the World Jewish Congress (WJC) and Vigo Social Intelligence collaborated to gather data on hate speech on social media, and specifically antisemitism.²¹⁰ In 2016, they identified 382,000 antisemitic posts on more than 100 platforms.²¹¹ The WJC and Vigo found that most of these posts attract little interest and do not go further: the

time of assessment compared to the results of the first monitoring exercise in December 2016" (*id.* at 3).

208 According to the EC's findings, "[d]ata shows a large disparity between IT companies when giving feedback to notifications made. While Facebook sent feedback in 93% of the cases, Twitter did so in only 32.8% of cases and YouTube in 20.7% of the cases. Twitter and YouTube provide more feedback when reporting comes from trusted flaggers" (*id.*).

209 European Commission, Results of Commission's Last Round of Monitoring of the Code of Conduct against Online Hate Speech, at <http://ec.europa.eu/newsro>.

210 The World Jewish Congress, in collaboration with Vigo Social Intelligence, *The Rise of Anti-Semitism on Social Media: Summary of 2016*. Vigo applied the IHRA criteria to public posts only (Facebook Messenger and WhatsApp are not included). Vigo divided online antisemitism into five categories: (1) expressions of hatred against Jews; (2) calls to harm Jews; (3) dehumanization of Jews; (4) Holocaust denial; (5) the use of symbols traditionally associated with antisemitism. Though this list does not include hate speech related to Israel, WJC and Vigo also show the relevant data on hatred for Israel (*id.* at 11-14).

211 *Id.* at 14.

average post is engaged by five surfers and has an average exposure of between 50 and 100 surfers. A total of 29 million surfers were exposed to antisemitic discourse in 2016. The WJC and VIGO also identified 3.3 million hate posts targeting Israel, Israelis, or the Israeli-Palestinian conflict. These were mainly about current political events and not spaced out equally over time.²¹²

The WJC and Vigo presented more detailed data in their report. For instance, 41% of the monitored antisemitic discourse included hate speech against Jews; 40% contained antisemitic symbols such as the swastika. In most cases (90%), the users who posted the hate speech did not come from groups of users identified as overtly antisemitic. The remaining posts included calls to harm Jews (8%), dehumanization (7%), and Holocaust denial (4%).²¹³ There were 31,000 posts urging attacks on Jews in 2016 (80 posts a day, or one every 20 minutes). Around 63% of all antisemitic discourse was found on Twitter, with the rest on blogs (16%), Facebook (11%), Instagram (6%), YouTube (2%), and other platforms (2%).²¹⁴ The WJC and Vigo also found that 68% of all online antisemitic discourse originated in the United States, followed by Germany (14%), the United Kingdom (4%), Canada (2%), and France (1.5%), with the rest from 30 additional countries. The WJC and Vigo concluded that racism and antisemitism have become normal.²¹⁵

A report issued in January 2018 shows an increase in daily (550) and hourly (23) posts that contain neo-Nazi and antisemitic symbols, as well as an increase in Holocaust denial. There was a decrease in antisemitic content on Facebook, Instagram, and YouTube, but an increase on Twitter

212 *Id.* at 15.

213 *Id.* at 14–17.

214 *Id.* at 39.

215 *Id.* at 15.

and blogs. In most countries, 2017 saw an increase in the number of posts using antisemitic symbols or denying the Holocaust compared to 2016. The United States leads the list, with a 36% increase in the use of antisemitic symbols and a 68% increase in Holocaust denial. Germany is the only country with a decrease in the use of neo-Nazi symbols (16% decrease), but not in Holocaust denial (2% increase).²¹⁶

Another organization that gathers information on antisemitic hate crime is the Anti-Defamation League (ADL). In its annual audit of antisemitic incidents, the ADL reported that, as a result of the 2016 presidential campaign in the United States, there was a massive increase in harassment of American Jews over 2015.²¹⁷ A more recent report, for the first nine months of 2017, indicated a rise of 67% in antisemitic incidents in the United States.²¹⁸ The political climate of the presidential campaign also led to the targeting of Jewish journalists. For the period August 2015 to July 2016, the ADL developed a set of keywords to capture antisemitic language on Twitter. Out of 2.6 million results, the ADL counted 19,253 overtly antisemitic tweets directed at 800 journalists.²¹⁹ These tweets

216 The World Jewish Congress, in collaboration with Vigo Social Intelligence, *Antisemitic Symbols and Holocaust Denial in Social Media Posts*, January 2018.

217 The surge occurred around the end of 2016 and the first three months of 2017. See ADL, *U.S. Antisemitic Incidents Spike 86 Percent So Far in 2017 after Surging Last Year, ADL Finds*.

218 "ADL Data Shows Anti-Semitic Incidents Continue Surge in 2017 Compared To 2016," ADL Israel (online).

219 One comment by the ADL is that the set of keywords is not inclusive, because it is impossible to predict all the "codes" used by antisemites to avoid censorship. Also, because many of the accounts have been deleted – whether by Twitter or their owners – the numbers presented are conservative. See ADL report, *Antisemitic Targeting of Journalists during the 2016 Presidential Campaign*, A Report from ADL's Task Force on Harassment and Journalism, October 19, 2016.

were viewed approximately 45 million times and sparked antisemitic content sent directly to journalists or other users. With this data, the ADL confirmed that the attacks were persistent and tended to come from self-identified nationalists and Trump supporters.²²⁰ According to the ADL, though many tweets were election-related, many others referenced classic antisemitic tropes.²²¹

Another method for tracking xenophobia and hate speech online employs content analysis, using conversation-analysis software such as Crimson Hexagon.²²² Pew Research Center used both content analysis and survey data to find that Americans are much more likely to view race-related posts than to post or share race-related content themselves—especially in the case of African Americans and Hispanics.²²³ Pew also found that an

220 The ADL found that 68% of the tweets were sent by 1,600 users (*id.*).

221 *E.g.*, Jews control the media, Jews control global finance, Jews perpetrated 9/11, etc.

222 "Crimson Hexagon is a software platform that identifies statistical patterns in words used in online texts. Researchers enter key terms using Boolean search logic so the software can identify relevant material to analyze. The Center draws its analysis sample from all public Twitter posts. Next, a researcher trains the software to classify documents using examples from those collected posts. Finally, the software classifies the rest of the online content according to the patterns derived during the training. Automated sentiment analysis, which is not perfect for analysis, had two stages: the first involves generating a list of terms to be included and excluded from the Boolean search; the second stage is training the algorithm to identify race-related tweets and to categorize them according to their subject matter." See Pew Research Center, August 2016, *Social Media Conversations about Race*.

223 68% of African American and 58% of Hispanic social-media users say that at least some of the posts they see on social networking sites are race-related. African Americans and Hispanics are also more likely to post or share content about race (*id.* at 5–8).

active race-related discussion on Twitter tends to follow social activism, such as the #BlackLivesMatter political and social movement.²²⁴

The Citizen Research Centre (CRC), too, has used Twitter to analyze the rise of online xenophobia. Looking at xenophobic posts on social media in South Africa from 2011 to 2017, it tracked incitement to violence and anti-immigrant content, nuanced opinions, and anti-xenophobia and anti-violence content.²²⁵ In South Africa, most of the conversation about xenophobia consists of shared news stories and international reports (e.g., refugees, Brexit, Trump), but other conversations were driven by individuals focusing on xenophobia in South Africa.²²⁶ At first, documented pro-xenophobia content accounted for only 1% of the conversations, but the figure rose to 4% in 2015 and 2016. Hateful anti-immigrant rhetoric increased in 2013 (16% of conversations) and reached a peak of 22% in 2014. But the CRC noted a decline during crises, suggesting “that [anti-immigrant rhetoric] is of more concern in building up to events than during the events themselves.”²²⁷ For anti-xenophobia, by contrast, the level of conversation remains low until a crisis emerges or an incident occurs and produces a substantial rise.²²⁸

Israel is no stranger to hate speech. The Berl Katznelson Foundation, in cooperation with Vigo Social Intelligence, created the Hate Speech Report, which tracks Hebrew-language hate speech in real time, including

224 *Id.* at 9–22.

225 Citizen Research Centre, *supra* note 2.

226 The CRC takes the entire public social-media conversation pertaining to xenophobia and looks only at content originating in South Africa. This enables it to segment the data into conversation themes and specific categories.

227 *Id.* at 19.

228 *Id.*

its sources and audiences.²²⁹ The report monitors online discourse for statements, phrases, and words that denote incitement, racism, exclusion, and violence. It also presents a detailed analysis of critical statements and events—for instance, how a statement by a public figure or an extreme event generated violent discourse in society.²³⁰ According to the report, from November 21, 2016, to November 20, 2017, there were more than five million racist expressions, curses, calls to violence, or offensive words—one every six seconds. Much of the hate speech targeted the media (a 500% leap within two years), but also government institutions including the president (up 220% within two years), the IDF Chief of Staff (up 500% within two years), and the Police Commissioner (up 60% within two years). Statements against the Israeli courts, including against specific judges, had risen by 230% within two years.²³¹

In summary, there are different methods for quantifying and tracking online racism and xenophobia. While state authorities usually stick to official criminal reports from the courts system and sometimes employ exercises, civil society relies on different methodologies, such as surveys and content analysis. The subjects monitored also vary. Some inquiries center on society at large, while others provide data on specific groups

229 According to the Hate Speech Report website (translated from Hebrew): "Vigo monitors more than half a million conversations every day on web portals, blogs, forums, public and private network and page responses, on a variety of social networks (Facebook, Twitter, Google+, YouTube, etc.). The data are segmented in real time by keywords and predefined parameters, which are embedded through an advanced technological system that has the ability to correct and learn. [...] The studies are conducted professionally and under full academic supervision, with an emphasis on analysis that enables the generation of operational insights into action (SWOT)." See [the Berl Katznelson Foundation's website \[in Hebrew\]; On Vigo Social Intelligence](#).

230 *Id.*

231 Berl Katznelson Foundation, *supra* note 2.

such as African Americans, Hispanics, Jews, and journalists. Finally, while most reviews look at incitement, as in South Africa, it is also possible to track anti-xenophobia and anti-violence content. Drawing on all types of data, mainly where the tracking employs the same methodology over time, can make it possible to propose policy solutions for combating hate speech and xenophobia. These solutions vary as a function of the context and of the actors who employ them. Before we enumerate the relevant actors and the policy instruments they use, it is essential that we understand the legal framework in which they work. Overall, despite the initial attempts to quantify and counter the phenomenon, online hate speech and xenophobia online are widespread and increasing.

Appendix B

Examples of Content Moderation by Several Major OSPs

Facebook

Facebook has an extensive content-moderation apparatus, but most of what is known about it comes from leaked documents and discussions with the policy managers. This system has been evolving ever since Facebook was incorporated and the platform developed.²³²

- **Statement:**

- Under “Safety,” Facebook’s Statement of Rights and Responsibilities (SRR) tells users that Facebook does its best to keep Facebook safe, but cannot guarantee it. “We need your help to keep Facebook safe, which includes the following commitments by you.” Among others, users “will not bully, intimidate, or harass any user.” Also, users “will not post content that: is hate speech, threatening, or pornographic; incites violence; or contains nudity or graphic or gratuitous violence.”
- Under “Protecting Other People’s Rights,” Facebook’s SRR tells users that they “will not post content or take any action on Facebook that infringes or violates someone else’s rights or otherwise violates the law.” Users cannot have names that are offensive or suggestive.²³³
- Facebook’s Community Standards state that “[w]e want people to feel safe when using Facebook. For that reason, we’ve developed a set of Community Standards, outlined below. These policies will help you understand what type of sharing is allowed on Facebook, and what type of content may be reported to us and removed.

232 Angwin and Grassegger, *supra* note 99.

233 See “What Names Are Allowed on Facebook,” [facebook.com](https://www.facebook.com/policy/terms).

Sometimes we will allow content if newsworthy, significant or important to the public interest—even if it might otherwise violate our standards. Because of the diversity of our global community, please keep in mind that something that may be disagreeable or disturbing to you may not violate our Community Standards.”

- On hate speech, Facebook’s Community Standards encourage respectful behavior. “People use Facebook to share their experiences and to raise awareness about issues that are important to them. This means that you may encounter opinions that are different from yours, which we believe can lead to important conversations about difficult topics. To help balance the needs, safety, and interests of a diverse community, however, we may remove certain kinds of sensitive content or limit the audience that sees it.

- **The Community Standards state further:**

- “Organizations and people dedicated to promoting hatred against these protected groups are not allowed a presence on Facebook.”

- “People can use Facebook to challenge ideas, institutions, and practices. Such discussion can promote debate and greater understanding. Sometimes people share content containing someone else’s hate speech for the purpose of raising awareness or educating others about that hate speech. When this is the case, we expect people to clearly indicate their purpose, which helps us better understand why they shared that content.”

- “We allow humor, satire, or social commentary related to these topics, and we believe that when people use their authentic identity, they are more responsible when they share this kind of commentary. For that reason, we ask that

Page owners associate their name and Facebook Profile with any content that is particularly cruel or insensitive, even if that content does not violate our policies. As always, we urge people to be conscious of their audience when sharing this type of content.”

■ “While we work hard to remove hate speech, we also give you tools to avoid distasteful or offensive content. Learn more about the tools we offer to control what you see. You can also use Facebook to speak up and educate the community around you. Counter-speech in the form of accurate information and alternative viewpoints can help create a safer and more respectful environment.”

○ The Community Standards refer to dangerous organizations, a category that includes organized hate groups.²³⁴ Facebook does not allow organizations or individuals that engage in terrorism or organized violence, or organized hate groups, to have a presence on Facebook.

○ According to the Community Standards, Facebook removes content that expresses support for groups that are involved in violent or criminal behavior. Supporting or praising leaders of these organizations, or condoning their violent activities, is not allowed. While Facebook “welcome[s] broad discussion and social commentary on these general subjects, [Facebook] ask[s] that people show sensitivity towards victims of violence and discrimination.”

○ With regard to public figures, Facebook does “permit open and critical discussion of people who are featured in the news or have a

234 Facebook, [Community Standards: Dangerous Individuals and Organizations](#).

large public audience based on their profession or chosen activities.”²³⁵ However, Facebook “remove[s] credible threats to public figures, as well as hate speech directed at them—just as we do for private individuals.”²³⁶ Content that appears to purposely target private individuals with the intention of degrading or shaming them will be removed.

- Finally, the Community Standards deal with content that mentions criminal activities or sexual violence and exploitation. In some situations, these might be indirectly relevant for determining what is hate speech.

- Material rule:

- Under its Community Standards, Facebook clarifies that it may remove hate speech. Under this rubric Facebook includes “content that directly attacks people based on their: race; ethnicity; national origin; religious affiliation; sexual orientation; sex, gender, or gender identity; or serious disabilities or diseases.

- Recently, ProPublica reviewed some of Facebook’s hate speech guidelines, which define how Facebook’s censors distinguish hate speech from legitimate political expression. According to ProPublica, Facebook has spent years developing these rules to distinguish between what should and should not be allowed on Facebook.²³⁷

235 Antigone Davis, *Protecting People from Bullying and Harassment*, FACEBOOK NEWSROOM (October 2, 2018).

236 Facebook’s Community Standards used to define private individuals as “people who have neither gained news attention nor the interest of the public, by way of their actions or public profession.”

237 In a recent talk with Prof. Jonathan Zittrain, Monika Bickert, Facebook’s head of global policy management, did not confirm whether these statements were still in force or if they have been updated. She did

- According to one guideline, Facebook deletes curses, slurs, calls for violence and other attacks only when they are directed at “protected categories.”²³⁸ For Facebook, this definition gives more leeway to users when they write about “subsets” of protected categories.
- According to ProPublica, for Facebook, a protected category plus an attack means hate speech, which content reviewers need to decide whether to delete or allow. For example, white men are a protected group because both traits (white and men) are protected. By contrast, female drivers and black children, like radicalized Muslims, are not protected subsets because one of their traits is not protected.
- There are also “quasi-protected” subgroups. For instance, migrants are protected only against calls for violence and dehumanizing generalizations. They are not protected against calls for exclusion or against degrading generalizations. According to ProPublica, the guidelines allow migrants to be referred to as “filthy,” but they cannot be likened to filth or disease—“when the comparison is in the noun form,” the document explains.
- According to ProPublica, there are some exceptions to the categories, as well as additional and more specific exemptions. For instance, there is a ban against advocating that anyone be sent to a concentration camp. However, because Nazis themselves are a hate group, the documents permit “Nazis should be sent to a concentration camp.”

mention that the report shows how much thought goes into the content-moderation process. See Berkman Klein Center for Internet & Society, *The Line between Hate and Debate on Facebook*, September 22, 2017).

238 As in the main material rule, these “protected categories” are based on race, sex, gender identity, religious affiliation, national origin, ethnicity, sexual orientation, and serious disability/disease.

- Facebook does not comply with the First Amendment's protection of free speech. According to Monika Bickert, Facebook's head of global policy management, its policies "do not always lead to perfect outcomes. That is the reality of having policies that apply to a global community where people around the world are going to have very different ideas about what is OK to share." Facebook's rule for itself is to allow free speech.
- Facebook's algorithm is designed to defend all races and genders equally. Here Facebook deviates from American law, which permits preferences such as affirmative action for racial minorities and women for the sake of diversity or redressing discrimination.
- Procedure:
 - Under "Protecting Other People's Rights," the Facebook SRR states that Facebook can remove any content or information posted by users if it believes that it violates the SRR or Facebook's policies.
 - The Community Standards state that Facebook removes content, disables accounts, and works with law enforcement when Facebook believe there is a genuine risk of physical harm or direct threats to public safety. When dealing with direct threats, Facebook notes that it "carefully review[s] reports of threatening language to identify serious threats of harm to public and personal safety. [It] remove[s] credible threats of physical harm to individuals." Facebook "may consider things like a person's public visibility or the likelihood of real-world violence in determining whether a threat is credible."
 - Under "Reporting Abuse," the Community Standards mention that Facebook's global community is growing every day, so it strives to welcome people to an environment free of abusive content. To do so, it relies on human beings; if users see something on Facebook that they believe violates its terms, they can report that content. It

has teams working around the world to review reported content. It explains, however, that a report of content does not guarantee that the content will be removed.²³⁹ For these situations, users can customize their experience.

- Facebook mentions that governments and law enforcement may ask it to remove content. Such requests may refer to content that violates local laws, even though it does not violate the Community Standards. After a careful legal review of the status of the content under local law, Facebook may make it unavailable only in the relevant country or territory.
- Facebook has guidelines for its content reviewers (human censors) on deleting posts. In May 2017, Mark Zuckerberg pledged to employ 7,500 content reviewers. They need to review the millions of reports Facebook receives every week.²⁴⁰ Reviewers need to make decisions within seconds and may vary in both interpretation and vigilance. Some of the guidelines tell content reviewers to take down posts by activists and journalists in disputed territories such as Palestine, Kashmir, Crimea, and Western Sahara. According to a report by the *Guardian*, reviewers may be underpaid and undervalued, receiving (at the time) roughly \$15 an hour.
- In addition, according to Monika Bickert, Facebook conducts weekly audits of every content reviewer's work. This is to ensure that Facebook's rules are being followed consistently.

239 Facebook, [What Happens When I Report Something to Facebook? Does the Person I Report Get Notified?](#)

240 See Mark Zuckerberg, www.facebook.com/zuck/posts/10103695315624661

- On highly political questions, Mark Zuckerberg intervenes in some cases and makes the final decision.²⁴¹ These may include a call by a political candidate to exclude protected groups.
- Facebook asks users to keep the following in mind:²⁴²
 - “[Facebook] may act anytime when something violates the Community Standards outlined here.
 - “Page owners may be asked to associate their name and Facebook Profile with a Page that contains cruel and insensitive content, even if that content does not violate our policies.
 - “Reporting something doesn’t guarantee that it will be removed because it may not violate our policies.
 - “Our content reviewers will look to reporting users for information about why a post may violate our policies. If you report content, please tell us why the content should be removed (e.g., is it nudity or hate speech?) so that we can send it to the right person for review.
 - “Our review decisions may occasionally change after receiving additional context about specific posts or after seeing new, violating content appearing on a Page or Facebook Profile.

241 Deepa Seetharaman, *Facebook Employees Pushed to Remove Trump's Posts as Hate Speech: Ruling by CEO Mark Zuckerberg to Keep Presidential Candidate's Posts Spurred Heated Internal Debates*, WALL STREET JOURNAL, October 21, 2016.

242 Facebook, [Community Standards: Hate Speech](#).

- “The number of reports does not impact whether something will be removed. Facebook never removes content simply because it has been reported more than one time.
- “The consequences for violating our Community Standards vary depending on the severity of the violation and the person’s history on Facebook. For instance, we may warn someone for a first violation, but if we continue to see further violations, we may restrict a person’s ability to post on Facebook or ban the person from Facebook.”
- Because not all disagreeable or disturbing content violates the Community Standards, Facebook enables users to customize and personalize their experience. Users can unfollow, or block or hide posts, people, pages, and applications they don’t want to see.²⁴³ Facebook then offers instructions on how to use “report links”:
 - First, users can use report links to send a message to the person who posted the content and request that the content be removed.
 - If users feel uncomfortable about reaching out to the speaker directly, Facebook suggests they reach out to a parent, teacher, or trusted friend, sharing the content and asking her or him to report the content to Facebook.
 - Facebook also makes it possible for users to block the instigator in question.

Facebook makes it possible for users to report different forms of problematic content, including profiles, posts, photos, videos, pages, groups, and events.²⁴⁴

How to Report Things

Don't have a Facebook account?

[Learn more](#) about how you can report potential abuse on Facebook.


The best way to report abusive content or spam on Facebook is by using the **Give feedback or report** link that appears near the content itself. To report a business you purchased something from on Facebook, you can fill out this [form](#).

Below are some examples of how you can report content to us:

Profiles

Posts

To report a post:

- 1 Click  in the top right of the post
- 2 Click **Report post** or **Report photo**
- 3 Select the option that best describes the issue and follow the on-screen instructions

Was this information helpful?

Yes No

[View Full Article](#)

[Share Article](#)

Posts on Your Timeline

Photos and Videos

Messages

Pages

Groups

244 Facebook, [Help Center: How to Report Things](#).

- Data:

Facebook has a dedicated website for government requests, both requests for data and requests to restrict access to content, based on local law. See <https://transparency.facebook.com/government/about/>.

Facebook also has a dedicated page for law-enforcement agencies, at www.facebook.com/safety/groups/law/guidelines.

According to Facebook, when governments submit content-related requests, Facebook studies the request to determine whether the content does indeed violate local laws. If Facebook determines that it does, the content is made unavailable in the relevant country or territory.

According to Facebook's data for January–June 2017, about 30 governments submitted content-related requests during that period. The leaders were Mexico (20,527), Germany (1,297), India (1,228), France (967), Turkey (712), Brazil (629), South Korea (572), Israel (472), Austria (363), and Italy (321).

Although these data are visible and accessible, Facebook does not create easily readable graphs, but only CSV files for downloading.

Google

Google has many services, but only one Terms of Service and privacy policy for most of them. Specific services, such as YouTube and Google Maps, have additional statements for the content shared on them.

- Statement:

- Google's Terms of Service say that Google services display some content that is not Google's. The content is the sole responsibility of the entity that makes it available. Google may review content to determine whether it is illegal or violates its policies, and Google may remove or refuse to display content that it reasonably believes

violates its policies or the law. But this does not necessarily mean that Google reviews content and one must not assume that Google does.

- According to the Terms of Service, automatic systems analyze users' content (including emails), but that is done to provide users with personalized, relevant product features such as customized search results, tailored advertising, and spam and malware detection. This analysis occurs as the content is sent and received and when it is stored. In other words, according to the Terms of Service, content is not checked or flagged for hate speech.

- Google also has a User Content and Conduct Policy for its social and sharing products and services. These products and services, according to Google, enable people from diverse backgrounds to start conversations, share experiences, collaborate on projects, and form new communities.²⁴⁵

- Google states that it depends heavily upon users' flagging of content that may violate its policies. After the flagging of a potential policy violation, Google may review the content and take action. This may be restricting access to the content, removing it, or limiting or terminating a user's access to Google's products. The decision may be affected by artistic, educational, or documentary considerations, or when there are other substantial benefits to the public from leaving the content as is.

- Specifically, for hate crimes, Google states that its products are platforms for free expression and that it does not support content that promotes hate speech. "This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line."²⁴⁶

245 Google, [Terms and Policies](#).

246 *Id.*

- In the case of terrorist content, Google does not permit terrorist organizations to use Google+ for any purpose. A user who posts content related to terrorism for educational, documentary, scientific, or artistic purposes must provide enough information for viewers to understand the context.²⁴⁷
- Google Maps is an example of a service with a specific policy regarding prohibited and restricted content. The policy applies to all formats, including reviews, photos, and videos. It does not allow content “that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics.” Google Maps does not accept content that is illegal or depicts illegal activity, including images of graphic or gratuitous violence, images that promote violence, or content produced by or on behalf of terrorist groups.
- YouTube also has a specific policy for its community. It asks users to show respect for other users’ trust. Google states that the community guidelines include “some common-sense rules that’ll help you steer clear of trouble.”²⁴⁸ It requests that YouTube users take these rules seriously. They are asked not to look for loopholes or to try to lawyer their way around the guidelines, but only to understand and respect them.
- YouTube repeats the Google definition of hate crimes. In addition, its policies state that “there is a fine line between what is and what is not considered to be hate speech. For instance, it is acceptable to criticize a nation state, but if the primary purpose of the content is to incite

247 *Id.*

248 YouTube, [Policies and Safety](#).


hatred against a group of people solely based on their ethnicity, or if the content promotes violence based on any of these core attributes, like religion, it violates our policy.”

- Material rule:
 - On YouTube, Google Maps, and other Google services, hate speech refers to content that “promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics.”²⁴⁹
 - “This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line.”
- Procedure:
 - Google Photos guides user on reporting content through the user interface:

If someone uses a shareable link to send you photos or videos that you believe violate [Google policies](#) you can report them.

Send a report

OPTION 1: Harassment, bullying, hate speech, graphic violence, sexually explicit content, or spam

1. Open the photo or video in Google Photos.
2. At the top right, select More , then **Report abuse**.
(If you don't see it, click **Sign in**. You need to be signed in to your Google Account to report something.)
3. Choose the reason for your report.
4. Select **REPORT**.

OPTION 2: Image of a minor

If you are the minor in the image, or if you are the parent or guardian of the minor, you can [request to restrict sharing of that image](#).

Actions we might take

After we receive your report, we may review the offending content and take action. Actions we might take:

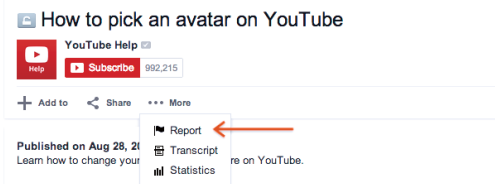
- Restrict access to the offending content
- Remove the offending content
- Limit or terminate a violator's access to Google products

Please keep in mind that something you think is offensive may not be spam or abuse according to [Google policies](#).

- For YouTube, on the other hand, Google states that its staff carefully reviews flagged content 24 hours a day, seven days a week, to determine whether there has been a violation of Google’s community guidelines. According to the YouTube Reporting Center, if no violations have been found, “no amount of flagging will change that, and the video will remain on [YouTube].”
- Flagging of videos is anonymous, so other users cannot tell who flagged a video.
- YouTube allows users to flag videos, thumbnails, comments, live chat messages, channels, and playlists.²⁵⁰

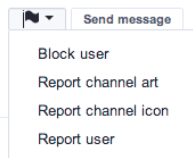
■ How to flag a video:

1. Sign in to YouTube.
2. Below the player for the video you want to report, click **More**.
3. In the drop-down menu, choose **Report**.
4. Select the reason that best fits the violation in the video.
5. Provide any additional details that may help the review team make their decision, including timestamps or descriptions of the violation.



■ How to flag a channel:

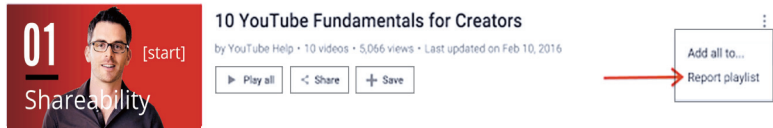
1. Sign in to YouTube.
2. Go to the channel page you want to report.
3. Click **About**.
4. Click the flag drop down.
5. Select the option that best suits your issue.



²⁵⁰ YouTube, [Help Center: Report Inappropriate Content](#).

■ How to flag a playlist:

1. Log in to Youtube
2. Go to the playlist content page you'd like to report
3. Click More
4. Select **Report Playlist**





- Google Maps allows users to flag content that violates Google Maps policies. Google's policy provides instructions on how to flag inappropriate content found on your listing or, alternatively, to fix your content that has been flagged or removed.²⁵¹ The policy asks users to flag only content that violates Google's policies and not content they simply don't like. Google also warns that it does not get involved in disputes between merchants.
- After inappropriate reviews that violate Google's policies have been flagged, the review will be assessed and possibly removed from the listing.

251 Google, [Flag and Fix Inappropriate Content](#).



Flag reviews in your account

If you find content that you believe violates our content policies, you can flag it for removal. The review will be assessed and possibly removed from your listing.

Computer

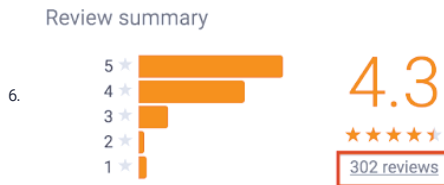
1. Sign in to [Google My Business](#).
2. If you have two or more listings, switch to card view  and click **Manage location** for the location you'd like to manage.
3. Click **Reviews** from the menu.
4. Find the review you'd like to flag, click the three dot menu , then click **Flag as inappropriate**.



Mobile

1. Open the Google My Business app.
2. Tap the menu , then tap **Reviews**.
3. Find the review you'd like to flag, tap the three dot menu , then tap **Flag review**.

Flag a review in Google Maps

1. Navigate to [Google Maps](#).
2. Search for your business using its name or address.
3. Select your business from the search results.
4. In the panel on the left, scroll to the "Review summary" section.
5. Under the average rating, click [number of] reviews.



7. Scroll to the review you'd like to flag, click the three dot menu , then click the flag icon .
8. Complete the form in the window that appears and click **Submit**.

- YouTube policies state there are two ways to report. Users can flag videos that violate YouTube's community guidelines. Users can also file an abuse report when multiple videos,

comments, or a user’s entire account are problematic. In these situations, a more detailed report must be submitted.²⁵²

- Google Maps also allows the flagging of photos, videos, questions, or answers. However, unlike regular reviews, the policy does not describe how Google acts after a user presses the “Submit” button.²⁵³
- YouTube also offers the following legal complaint form:

Hate Speech Legal Complaint

⚠ Please note that abuse of our legal forms may result in the termination of your YouTube account.

If a video contains your personal information without consent, including your image, name or national identification number, please contact us through our [Privacy Complaint Process](#).

Country of complaint:^{*}

Your full legal name (aliases, usernames or initials are not accepted):^{*}

Your email address: XXXXXXXXXX

Source of allegedly illegal content:^{*}

⚠ If you wish to report potential copyright infringement, please submit a formal [Copyright Infringement Notification](#). Alleged copyright violations should not be submitted via this contact form.

Please cite the specific hate speech law that the content is allegedly violating:^{*}

Please provide a hyperlink to the text of that law:

Please select the type of infringing content that you are reporting:^{*}

Please select one: ▼

YouTube Content URL:^{*}

Add another URL

Please describe how the content in question is allegedly violating the law, including the specific statements and/or images that allegedly violate the hate speech law that you’ve cited (if applicable, include a timestamp).^{*}

Agree to the following statement:^{*}

I declare that the information in this notice is true and complete.

Typing your full name in the box below will act as your digital signature ^{*}

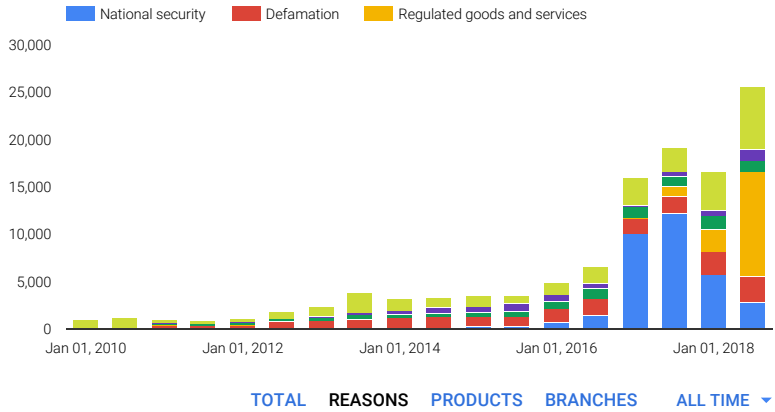
Please note that data will be stored in accordance with Google’s [privacy policy](#).

By submitting a legal complaint to YouTube, you agree that YouTube may send the legal complaint to the [Lumen project](#). Before publicly posting a notice of your submission, the Lumen project will redact any personal contact information (e.g., phone number, email, physical address) and we may publicly share the link to that complaint. You can see an example of a Lumen notice [here](#)

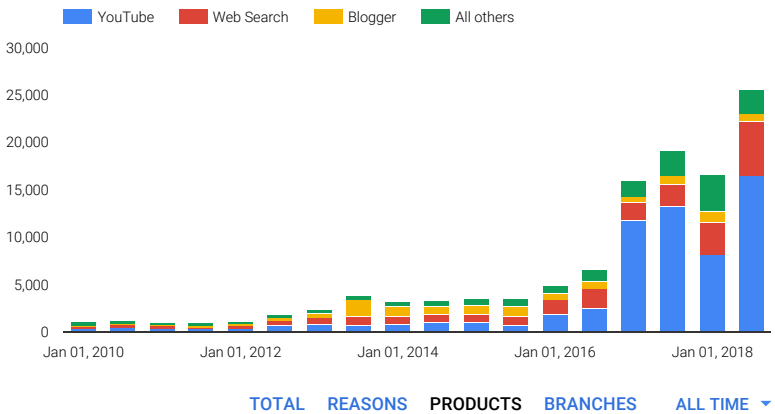
252 YouTube, [Hate Speech Policy](#).

253 Google, [Flag and Fix Inappropriate Content](#).

- Data:
 - Google government reports can be found at <https://transparencyreport.google.com/>.
 - According to Google, it receives content-removal requests through a variety of channels and from all levels and branches of government—court orders, written requests by national and local government agencies, and requests by law-enforcement professionals. Google receives complaints from government bodies and courts that content violates local laws; these are often not directed at Google. Sometimes users will forward government removal requests to Google, such as when a person attaches a court order declaring certain content to be illegal. Some requests ask for the removal of multiple content items; conversely, there may be multiple requests for the removal of the same item.
 - Google requires court orders rather than government requests. It examines the legitimacy of every document and notes that some government requests have been falsified.
 - Google always evaluates requests. They must be in writing, be as specific as possible about the content to be removed, and clearly explain how the content is illegal. Google does not honor requests that have not been made through the appropriate channels.
 - Google has an interactive website that allows viewers to learn about requests based on the total number of requests, the reasons for the requests, the relevant products, and more. The data goes back to 2009.
 - Reasons for government requests categorized based on reasons for content removal:

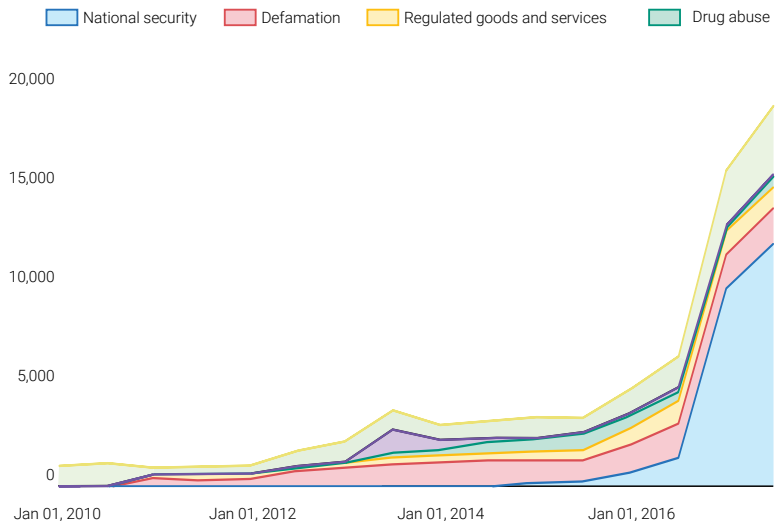


○ Google also displays the reasons for government requests, categorized by products:

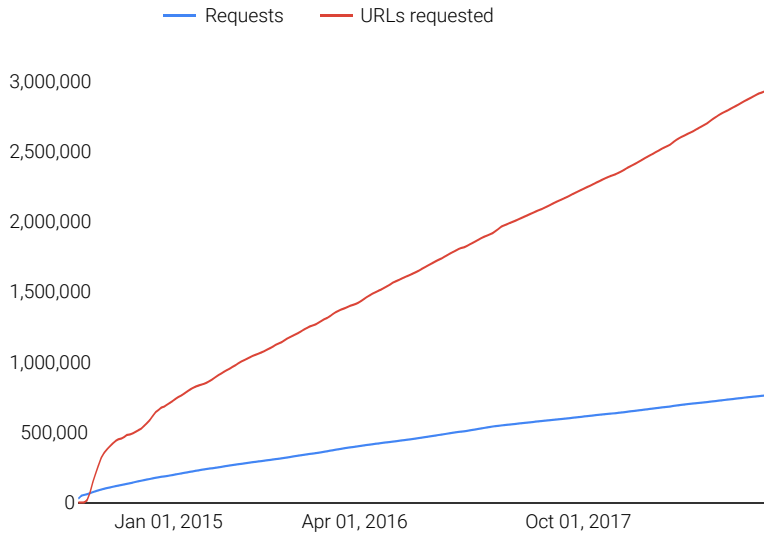


○ Google counts the reasons why government ask for content removal. These data go back to December 2010:

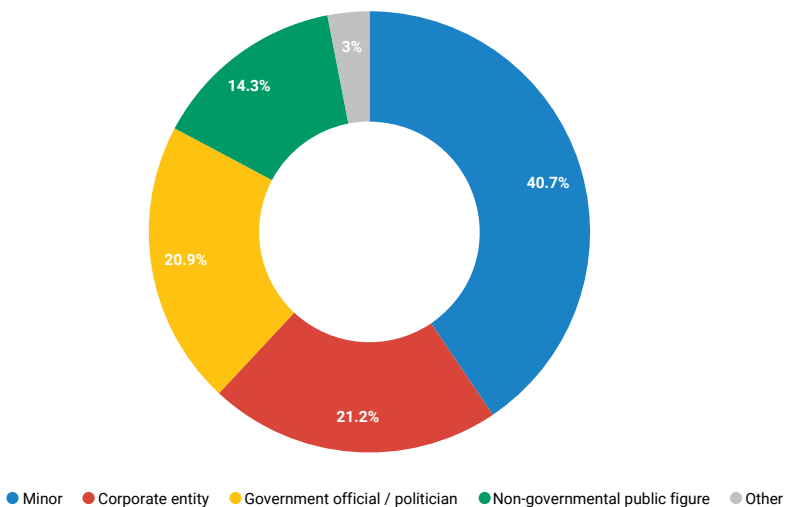
Reasons cited for content removal



- Google provides data on delist requests based on the European “right to be forgotten,” the court-ordered right that allows users to ask search engines to remove certain results from queries. The search engine must comply if the links are “inadequate, irrelevant or no longer relevant, or excessive.” The search engine needs to take into account public interest factors, such as if the individual is a public figure.
- According to Google, it delists only the URL associated with the person’s name and only from Google’s European search results, but not for the rest of the world. Since May 29, 2014, Google received more than 790,103 requests with more than 3 million URLs to be delisted. According to Google, it decided not to delist in 55.7% of these cases.



○ According to Google, about 88% of the requests were made by private individuals. The other requests were associated with minors, corporations, government officials or politicians, non-governmental public figures, and others:



Microsoft

- Statements:
 - Microsoft's Terms of Use have little to say about the use of Microsoft services. These services may include e-mail, bulletin boards, chat areas, news groups, forums, communities, personal web pages, calendars, photo albums, file cabinets, and/or other message or communication facilities designed to enable users to communicate with others.
 - Microsoft's Terms of Use mandate that users will not:
 - Defame, abuse, harass, stalk, threaten, or otherwise violate the legal rights (such as rights of privacy and publicity) of others;
 - Publish, post, upload, distribute, or disseminate any inappropriate, profane, defamatory, obscene, indecent or unlawful topic, name, material or information;
 - Violate any applicable laws or regulations;
 - Violate any code of conduct or other guidelines which may be applicable for any particular Communication Service.²⁵⁴
 - Microsoft also has the Microsoft Services Agreement, which applies to services such as Bing, Cortana, Microsoft Accounts, Office, OneDrive, Windows Store, and Xbox.
 - The Microsoft Services Agreement includes a section titled "Code of Conduct" (i.e., not a separate document). In this section, users agree that they:
 - Will not do anything illegal;

254 [Microsoft Terms of Use](#) (last updated: June 24, 2015).

- Will not publicly display or use the services to share inappropriate content or material (involving, for example, nudity, bestiality, pornography, graphic violence, or criminal activity);
 - Will not engage in activity that is harmful to themselves, the services, or others (e.g., transmitting viruses, stalking, posting terrorist content, communicating hate speech, or advocating violence against others);
 - Will not help others break these rules.
- In the Services Agreement, Microsoft enumerates its enforcement rights:
- “If [users] violate these Terms, [Microsoft] may stop providing services to [users] or [Microsoft] may close [users’] Microsoft account or Skype account.”
 - “[Microsoft] may also block delivery of a communication (like email or instant message) to or from the services in an effort to enforce these terms or [Microsoft] may remove or refuse to publish [users’] content for any reason.”
 - “When investigating alleged violations of these terms, Microsoft reserves the right to review [users’] content in order to resolve the issue. However, [Microsoft] cannot monitor the entire services and make[s] no attempt to do so.”²⁵⁵
- Microsoft has a more detailed code of conduct for Xbox Live. It explains “what conduct is” and what conduct Microsoft prohibits. Conduct is anything you do that impacts yourself, others, Microsoft, or Xbox Live. Microsoft provides examples of conduct that is not permitted:

- Do not create, share, use, or promote prohibited content.
- Do not engage in illegal activity. For example, do not threaten to hurt others physically; spread lies about someone, a product, a business, or a group.
- Do not harm or harass. For example, do not encourage violence against people or animals; or scream at, intimidate, or bully others.
- The Xbox Live code of conduct also explains what content is and which content is prohibited. “Content is anything you create, share, use, or promote that another person could see or hear or otherwise experience, like Gamertags, profile information, in-game content, and videos.
 - Content that involves illegality, e.g., terrorism or criminal activities, is prohibited.
 - Content that could harm or harass a person, including oneself, or an animal. For instance, negative speech (including hate speech or threats of harm) directed at people who belong to a group, including groups based on race, ethnicity, nationality, language, gender, age, disability, veteran status, religion, or sexual orientation/expression.
- Material rule:
 - The Xbox code of conduct defines negative speech, which includes hate speech. Negative speech is speech that is directed at people who belong to a group, including groups based on race, ethnicity, nationality, language, gender, age, disability, veteran status, religion, or sexual orientation/expression.
- Data:
 - Microsoft publishes content removal requests on its corporate responsibility page, located at www.microsoft.com/en-us/about/corporate-responsibility/crrr/.

- According to Microsoft, when it receives a government request to remove content it carefully reviews and assesses:
 - The request, in order to understand the reason for it
 - The requesting party's authority
 - The applicable policies or terms of use for the affected product or service
 - Microsoft's commitments to its customers and users with regard to freedom of expression.

Based on this review, Microsoft determines whether and to what extent it should remove the content in question. The report includes government requests for the removal of content for Microsoft online consumer services, such as Bing, OneDrive, Bing Ads, and MSN.

- According to Microsoft, between January and June 2018 it received 732 requests to remove content, from eleven governments: Australia, China, France, Germany, Israel, Kazakhstan, the Netherlands, Russia, South Korea, Taiwan, and the United Kingdom.²⁵⁶
- Microsoft took action on 586 of the 732 requests (80%). Of the 39 requests to close an account, Microsoft acted on 20 (51%).²⁵⁷
- Microsoft also received "right to be forgotten requests." In January–June 2018, Microsoft received and processed 2,780 requests for 9,132 URLs. Microsoft accepted 5,043 requests (55%). Overall, between May 2014 and June 30, 2018, Microsoft received and processed 26,729 requests for 78,781 URLs. It accepted 32,725 of them (42%).²⁵⁸
- Microsoft also makes its revenge porn removal requests available. Between January to June 2018, Microsoft received 362 request reports, of which it accepted 242 (67%).²⁵⁹

²⁵⁶ Microsoft, [Content Removal Requests Report](#).

²⁵⁷ *Id.*

²⁵⁸ *Id.*

²⁵⁹ *Id.*

Twitter

Twitter's Terms of Service differentiate between US-based consumers and those located outside the United States. The Twitter Rules are similar across the globe.

- Statements:
 - According to Twitter's Terms of Services, users "are responsible for [their] use of the Services and for any Content [they] provide, including compliance with applicable laws, rules, and regulations."
 - Twitter's Terms of Service tells users to understand that by using Twitter's services they may be exposed to content that might be offensive, harmful, inaccurate, or otherwise inappropriate, or in some cases, posts that have been mislabeled or are otherwise deceptive. The content is the sole responsibility of the person who originates such content.
 - Twitter does not endorse, support, represent, or guarantee the completeness, truthfulness, accuracy, or reliability of any content or communications posted via the services or endorse any opinions expressed via the services.
 - Interestingly, Twitter differentiates between American and non-American users. Twitter tells US-based consumers that it reserves the right to remove content alleged to be a violation or infringement without prior notice, at its sole discretion, and without liability vis-à-vis users. Outside the United States this statement is broader: Twitter reserves the right to remove content that violates its terms, including unlawful conduct and harassment.
 - Twitter does not monitor or control content posted via its services and cannot take responsibility for such content. However, Twitter may remove or refuse to distribute any content on its services, suspend or terminate user accounts, and reclaim usernames without liability vis-à-vis users.

- Twitter asks users to review the Twitter Rules, which are part of the user agreement (alongside Twitter’s privacy policy and terms of service). The Twitter Rules outline what is prohibited on Twitter’s services. Users may use Twitter’s services only in compliance with these terms and all applicable laws, rules, and regulations.
- The Twitter Rules start by mentioning the enforcement actions that Twitter can take for failure to adhere to the policies. These enforcement actions include:
 - (1) Requiring users to delete prohibited content before they can create a new post or interact with other users;
 - (2) Temporarily limiting users’ ability to create posts or interact with users;
 - (3) Asking users to verify their account ownership using their phone or email;
 - (4) Permanently suspending users’ existing and future account(s).
- In addition, the Twitter Rules include two specific statements about hateful conduct and imagery:
 - “Hateful conduct: [Users] may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.”
 - “Hateful imagery and display names: [Users] may not use hateful images or symbols in [their] profile image or profile header. [Users] also may not use [their] username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.” Enforcement of this rule began on December 18, 2017.

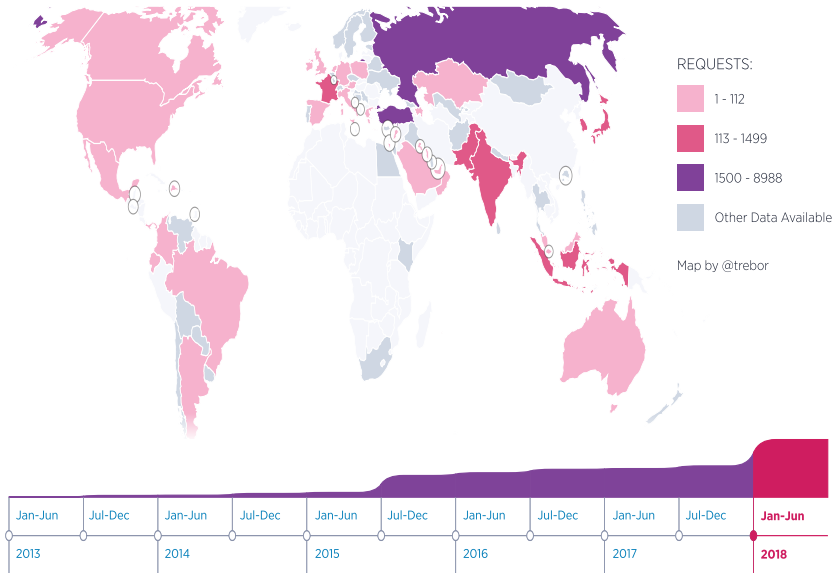
- With regard to the Twitter Rules, Twitter summarizes, stating that “accounts under investigation or which have been detected as sharing content in violation of these Rules may have their account or Tweet visibility limited in various parts of Twitter, including search.”
 - Twitter also has a “hateful conduct policy”: “Freedom of expression means little if voices are silenced because people are afraid to speak up. We do not tolerate behavior that harasses, intimidates, or uses fear to silence another person’s voice. If you see something on Twitter that violates these rules, please report it to us.”
 - Finally, Twitter has rules for users who automate their activity on Twitter. Twitter clarifies that automated activity is subject to the Twitter Rules and that users should carefully review the policies to ensure that their automated activities are compliant. Automated applications or activities that violate these policies or that facilitate or induce users to violate them may be subject to enforcement action, potentially including suspension of associated Twitter accounts. The automation rules apply, *inter alia*, to automated abusive behavior that encourages, promotes, or incites abuse, violence, hateful conduct, or harassment, on or off Twitter.
- Material rule:
 - “To ensure that people feel safe expressing diverse opinions and beliefs, [Twitter] prohibits behavior that crosses the line into abuse, including behavior that harasses, intimidates, or uses fear to silence another user’s voice. The context matters when evaluating for abusive behavior and determining appropriate enforcement actions. Factors Twitter may take into consideration include but are not limited to whether:

- “the behavior is targeted at an individual or group of people;
 - “the report has been filed by the target of the abuse or a bystander;
 - “the behavior is newsworthy and in the legitimate public interest.”
- Both the Twitter Rules and Twitter’s hateful conduct policy explain that users may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
- Twitter gives examples of what it does not tolerate under the hateful conduct policy. These include behavior that harasses individuals or groups of people with:
 - violent threats;
 - wishes for the physical harm, death, or disease of individuals or groups;
 - references to mass murder, violent events, or specific means of violence in/with which such groups have been the primary targets or victims;
 - behavior that incites fear about a protected group;
 - repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone.
- Procedures:
 - Twitter lists some of its enforcement mechanisms:
 - Context matters. Some Tweets may seem to be abusive when viewed in isolation but may not be when viewed in the context of a larger conversation. While Twitter accepts reports of violations from anyone, sometimes it also needs to hear

directly from the target to ensure that it has a proper context. In addition, the number of reports that Twitter receives does not impact whether or not something will be removed. However, it may help Twitter prioritize the order in which it gets reviewed.

- Twitter focuses on behavior. Twitter enforces policies when someone reports behavior that is abusive and targets an entire protected group and/or individuals who may be members. This targeting can happen in any manner (for example, @mentions, tagging a photo, and more).
 - The consequences of violating the rules vary depending on the severity of the violation and the person's previous record of violations. Twitter may ask users to remove an offending Tweet before they can Tweet again. Twitter may also suspend an account.
- Data:
 - According to Twitter, the removal requests it receives are generally about content that may be illegal in a specific jurisdiction. Governments (including law-enforcement agencies), organizations chartered to combat discrimination, and lawyers representing individuals are among the complainants. The data presented below refer only to official requests.
 - Twitter's website shows an interactive map of the requests it received.

Removal Requests Worldwide



- For instance, Twitter explained that between January and June 2017 global removal requests affected a total of 14,120 accounts, as follows: 1,760 accounts had some content withheld (account-level or tweet-level); 3,023 had some content removed for violating Twitter’s Terms of Service. No action was taken on the remaining requests (9,337).
- According to Twitter, roughly 90% of the removal requests between January and June 2017 originated from only four countries: France, Germany, Russia, and Turkey. Turkey submitted the most requests, accounting for approximately 45% of the worldwide total.
- From January to June 2017, Twitter received eight requests to remove content from verified Twitter accounts of journalists or

news outlets. Twitter did not act on any of these requests because of their political and journalistic nature.

- In addition, during this reporting period, Twitter received 1,336 requests from Twitter’s external “trusted reporters.” These are organizations that have a mandate to report content that may be considered hate speech under local European laws and which have entered into a formal partnership with Twitter.
- Twitter also has a Country Withheld Content (CWC) tool. Since 2012, Twitter has applied the CWC tool in 13 countries: Australia, Brazil, France, Germany, Great Britain, India, Ireland, Israel, Japan, the Netherlands, Russia, Spain, and Turkey. From January to June 2017, Twitter withheld content at the account or tweet level in 10 of those 13 countries (except India, Ireland, and Israel).

GoDaddy.com

- Statements:

- According to the GoDaddy.com terms of service, “you” (unspecified) will not use the site and its services in a “manner” that is, among others:

- illegal, or promotes or encourages illegal activity;
- promotes, encourages or engages in terrorism, violence against people, animals, or property.

The definition of “in a manner” is left to GoDaddy’s sole and absolute discretion.

- According to GoDaddy, it does not pre-screen user content posted to a website hosted by GoDaddy.com or posted on its site. However, GoDaddy reserves the right but undertakes no duty to perform pre-screening. GoDaddy can decide whether any item of user content is appropriate and/or complies with GoDaddy.com policies.

- GoDaddy also “expressly reserves the right to deny, cancel, terminate, suspend, lock, or modify access to (or control of) any Account or Services (including the right to cancel or transfer any domain name registration) for any reason (as determined by GoDaddy in its sole and absolute discretion), including but not limited to the following.” Among others:
 - to comply with court orders or subpoenas;
 - “to avoid any civil or criminal liability on the part of GoDaddy, its officers, directors, employees and agents, as well as GoDaddy’s affiliates, including, but not limited to, instances where [users] have sued or threatened to sue GoDaddy”;
 - “to respond to an excessive amount of complaints related in any way to your Account, domain name(s), or content on your website.”
- “GoDaddy, its officers, directors, employees, agents, and third-party service providers shall not be liable to you or any other person or entity for any direct, indirect, incidental, special, punitive, or consequential damages whatsoever, including any that may result from, among others: ... Any user content or content that is defamatory, harassing, abusive, harmful to minors or any protected class, pornographic, ‘x-rated,’ obscene or otherwise objectionable.”
- Procedures:
 - In a report on inappropriate content of disturbing imagery, violence, etc., visitors must attach the relevant URL and write a short explanation with details or explanations about why they are reporting or how the content is offensive.

Inappropriate Content

1**Report the Details****2**

Confirmations

Details/Explanation:

- According to the GoDaddy.com terms of service, GoDaddy may:
 - “Remove any item of User Content and/or terminate a User’s access to its site or services on its site for posting or publishing any material in violation of GoDaddy’s policies (as determined by GoDaddy in its sole and absolute discretion), at any time and without prior notice.
 - “Terminate a User’s access to its site or services on its site if GoDaddy has reason to believe the User is a repeat offender.
 - “If GoDaddy terminates access to its site or services on its site, GoDaddy may, in its sole and absolute discretion, remove and destroy any data and files stored by you on its servers.”

Appendix C

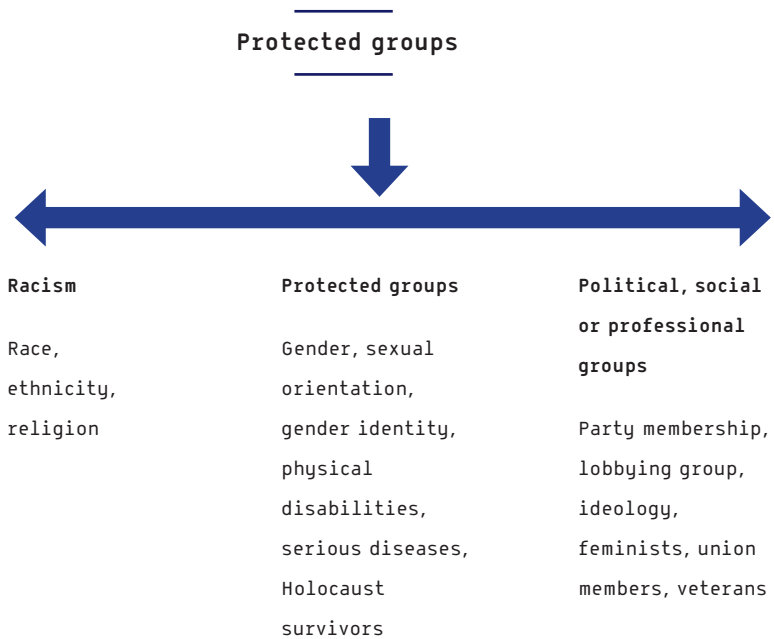
Applying the Proposed Model to Twitter's Community Standards

What follows implements and analyzes the material criteria for Twitter. To conduct the analysis we used the policy rules identified in the Twitter Rules that aim to protect Twitter users' experience and safety.

1. The speech targets a group or an individual as a member of a group: Currently, Twitter prohibits users from promoting violence against, threatening, or harassing other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.²⁶⁰ According to reports from September 2018, Twitter will prohibit "content that dehumanizes others based on their membership in an identifiable group, even when the material does not include a direct target."²⁶¹ As such, Twitter is located at the middle of the continuum.

²⁶⁰ Twitter, *Hateful Conduct Policy*.

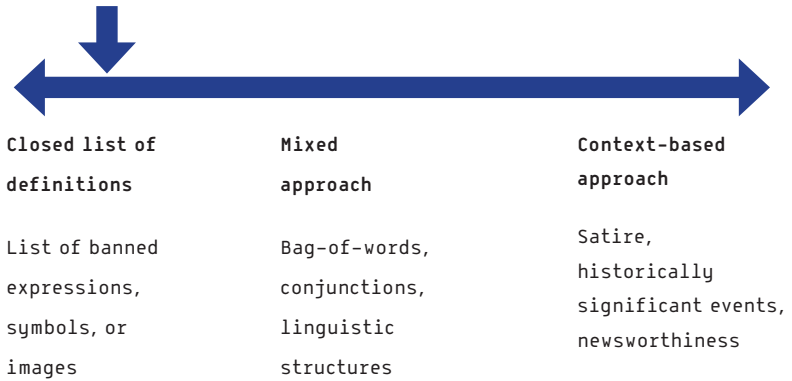
²⁶¹ Louise Matsakis, *Twitter Releases New Policy on "Dehumanizing Speech,"* WIRED, September 25, 2018. According to this report, the new policy expands "upon Twitter's existing hateful conduct policies prohibiting users from threatening violence or directly attacking a specific individual on the basis of characteristics such as race, sexual orientation, or gender."



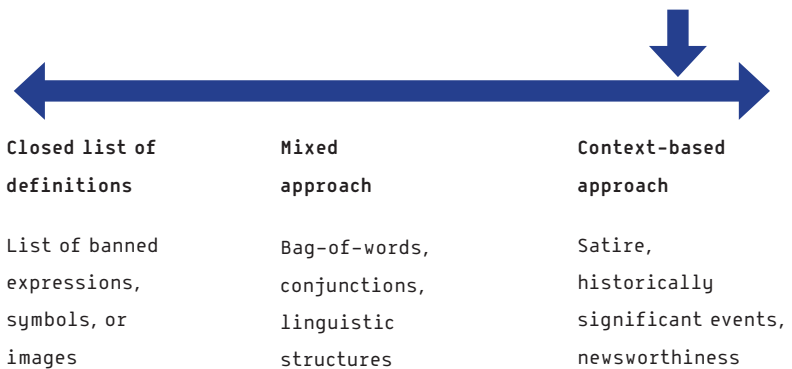
2. **The speech expresses hatred:** According to the Twitter Rules, users are not allowed to use hateful imagery and symbols in their profile image or profile header.²⁶² Users are also not allowed to use their username or display name to engage in abusive behavior. Hence, for display names, Twitter's policy is located under the closed list of definitions.

262 Twitter, *The Twitter Rules*.

Definitions of expressions of hatred
(from closed list to context-based):

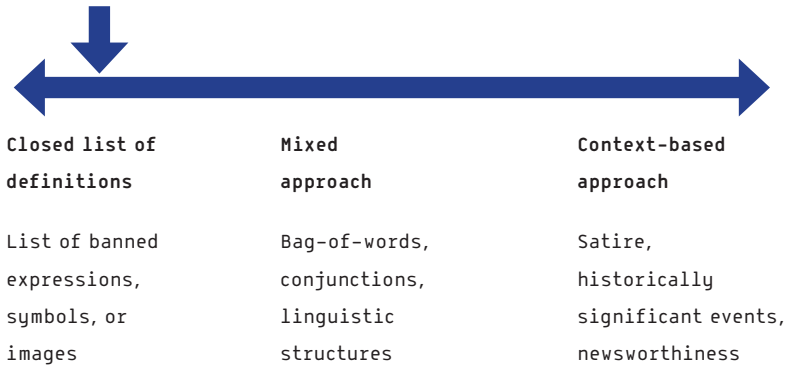


In contrast, Twitter sets a different rule for posts. According to the Twitter Rules, context matters when it evaluates whether behavior is abusive and determines appropriate enforcement actions. For Twitter, some tweets may seem abusive when viewed in isolation but not when viewed in the context of a larger conversation. Twitter takes into consideration whether the behavior is targeted at an individual or a group of people.²⁶³

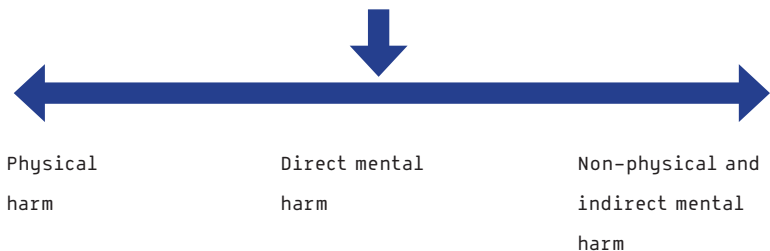


²⁶³ Twitter, *Hateful Conduct Policy*.

At the same time, the Twitter rules also state that it does not tolerate references to mass murder or violent events in which such groups have been the primary targets or victims.²⁶⁴



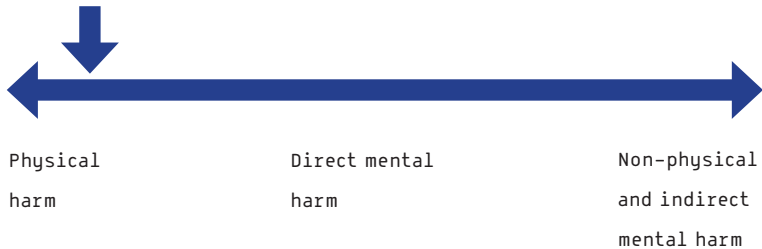
3. The speech could cause harm to an individual: To ensure that users feel safe expressing diverse opinions and beliefs, Twitter prohibits behavior that crosses the line into abuse. Abuse, according to Twitter, includes behavior that harasses, intimidates, or employs fear to silence another user's voice.²⁶⁵ Twitter also does not tolerate behavior that incites fear about a protected group. Hence, Twitter's policy deals with "direct mental harm."



²⁶⁴ Twitter, *Violent Threats and Glorification of Violence*.

²⁶⁵ Twitter, *The Twitter Rules*.

In addition, under its hateful conduct policy Twitter offers examples of what it does not tolerate. This includes violent threats and a desire for physical harm to or the death or illness of individuals and groups.²⁶⁶

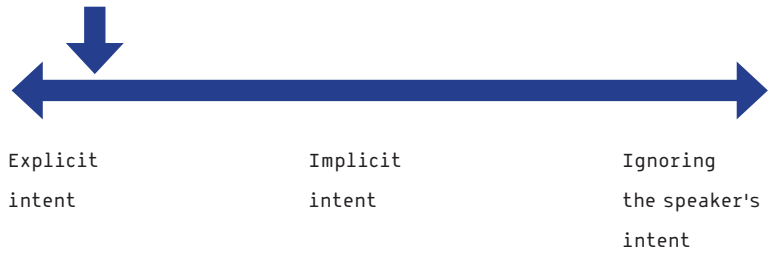


4. **The speaker intends to harm:** The Twitter Rules mention intent in specific cases.²⁶⁷ According to its rules on violent threats and glorification of violence, Twitter considers “threats to be explicit statements of one’s intent to kill or inflict serious physical harm against another person. This includes, but is not limited to, threatening to murder someone, sexually assault someone, break someone’s bones, and/or commit any other violent act that may result in someone’s death or serious injury.” Vague threats, on the other hand, and wishing or hoping that someone experience serious physical harm or threatening less serious forms of physical harm do not fall under the violent threat policies and may be reviewed under the abusive behavior and hateful conduct policies.²⁶⁸

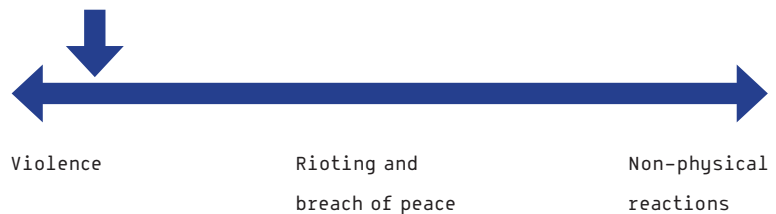
266 Twitter, *Hateful Conduct Policy*.

267 According to Sellars, *supra* note 27, Twitter used to address underlying intent in its policies against conduct that promotes violence or directly attacks a group.

268 Twitter, *Violent Threats and Glorification of Violence*.

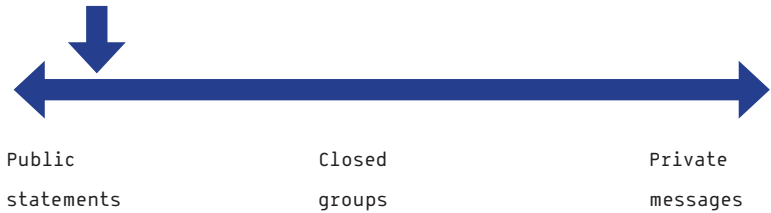


5. **The speech incites to socially undesirable action:** According to Twitter, the rationale behind its policy on violent threats and the glorification of violence is that the company wants “Twitter to be a place where people feel safe to freely express themselves. Thus, [Twitter] will not tolerate behavior that encourages or incites violence against a specific person or group of people. [Twitter] also takes action against content that glorifies acts of violence in a manner that may inspire others to replicate those violent acts and cause real offline danger, or where people were targeted because of their potential membership in a protected category.”²⁶⁹



269 *Id.*

6. **The speech is public:** Twitter sets all messages as public and thus deals only with public statements. While users can send each other private messages, Twitter does not have an interface that supports closed group discussions. At the same time, Twitter does take into consideration whether the behavior is targeted at an individual or a group of people.



Conclusions from the Twitter case-study:

Based on the foregoing analysis, Twitter employs a very strict approach that is willing to delete content that appears to be hate speech. Nevertheless, Twitter's policy is very broad and hard to define. In some cases, the policy treats the same issue in different ways. For instance, if a header is deemed offensive, the entire user should be deleted. Twitter must create a unified and clearer rule, one that is less offensive and intrusive regarding user headers and usernames, and less intrusive regarding regular content.

Realigning the Law to Better Uphold the State's Duty to Protect Human Rights

Towards an Interoperable Model for Addressing Racism and Strengthening Democratic Legitimacy

Karen Eltis | Ilia Siatitsa

Abstract | Introduction | PART I: Identifying Hate Speech | PART II: New Directions for Public and Private Accountability

ABSTRACT

Regulation of online hate speech is one of the thorniest issues that confront us in the digital age. First, the context-specific character of hate speech renders it dependent on specific circumstances at a particular

moment in time. This is intensified by the online space's potentially timeless and cross-cultural reach. Second, online spaces often amplify the nefarious effects of certain otherwise isolated statements that in the analog world would not have had the same potentially devastating effect. The discussions thus far have focused primarily on pushing the platforms that host the problematic content to remove it as soon as it is flagged or identified. This paper engages in a comparative analysis of national approaches to regulating hate speech and of the human rights approach to hate speech, in an effort to provide greater definitional clarity. It accordingly recommends the adoption of universal "hate speech" indicators informed by human rights law practices; interim measures by platforms and other relevant actors to mitigate amplification harms; and the introduction of a tiered approach to oversight to ensure a reasoned decision and allow private censorship decisions to be adequately reviewed.

Introduction

This report delves into the diverse legal responses to restrictions on speech (both online and offline) in selected parts of the world and endeavors to highlight how internet communications challenge traditional normative assumptions. The recommendations that stem from this analysis aim at the eventual development of an interoperable model¹ for addressing racism online, in order to satisfy the state's commitment to safeguarding human rights and to effectively but fairly allocate supervisory burdens to all actors involved.

A. Curtailing Chilling Speech—Ultimately a Matter of Protecting Democracy, Equality, and Free Expression Itself

Hate speech,² as Jeremy Waldron powerfully argues, should be regulated “as part of our commitment to human dignity and to inclusion and respect for members of vulnerable minorities.” Otherwise, society's most persecuted members risk being further silenced, both individually and collectively, and alienated from the larger political community. Ironically,

1 The term "interoperable" emphasizes the importance of re-conceptualizing digital privacy in a more trans-systemically viable fashion. It may ultimately help overcome cultural barriers for the purpose of a transnational legal exchange or forging a "conceptual middle ground." See Karen Eltis, *The Privacy Divide: Bridging the Gap between Legal Traditions*, CBA NATIONAL, January 12, 2017, www.nationalmagazine.ca.

2 One of the problems of hate speech is the lack of agreement regarding its meaning. The definition by Ruth Wedgwood is one example: Ruth Wedgwood, *Freedom of Expression and Racist Speech*, 8 TEL AVIV U. STUD. L. 325 (1988).

then, more speech in the networked context does not shepherd us to truth, but may operate to silence vulnerable groups. Such alienation is further amplified in the digital age, when misinformation, as we now know, spreads considerably and disproportionately faster on digital platforms. These platforms by their nature seek to maximize engagement,³ effectively frustrating “the very meaning of intelligent democracy [as] the ability to contemplate all alternatives and talk about them rationally.”⁴ Accordingly, the medium allows for unprecedented dissemination of propaganda, emboldening of racism, disruption of elections, and distortion of civil engagement.

Racist vitriol quickly overshadows facts,⁵ effectively paving the way for violence, as the ethnic cleansing in Myanmar so shockingly confirms. As Alexander Bickel forcefully cautioned, “where nothing is unspeakable, nothing is undoable.”⁶ Whereas the ultimate futility of speech regulation has in many cases correctly deterred the United States and Canada from going that route, the reality today is that online expression is already being heavily censored, albeit by private actors, often at the informal behest of governments or undemocratic regimes (rather than through the law or by virtue of express authorization to do so). It is therefore no longer a question of whether to limit speech on the internet (because arbitrary and unevenly applied limits are already firmly in place), but of who will limit it and by what authority, and whether democratic countries and institutions, including an independent judiciary (which reviews actions

3 See, e.g., Amy Davidson Sorskin, *The Age of Donald Trump and Pizzagate*, THE NEW YORKER, December 5, 2016.

4 Wedgwood, *supra* note 2, at p. 328.

5 Soroush Vosoughi, Deb Roy, & Sinan Aral, *The Spread of True and False News Online*, 6380 SCIENCE 1146 (2018).

6 ALEXANDER BICKEL, THE MORALITY OF CONSENT 73 (1975).

based on a principled framework), will be a part of that conversation, guided by the substantive rule of law rather than by backroom deals with reluctant corporate actors.

More broadly still, policymakers must recognize that artificial intelligence (AI) is being deployed as a decision-making tool, as part of the more general effort to do away with intermediaries.⁷ In international law, the responsibility to uphold human rights is seen as a central justification for state institutions to exercise power. The issue of how that translates to multinational corporations and their regulation of expression is now of the essence. Indeed, under well-established doctrines, states cannot relinquish their duty to protect human rights, even in cooperation with actors that function in a manner akin to governments. Beyond the question of the compatibility of such practices with the states' obligations to protect freedom of expression, the automatization of content control and removal seems to be at odds with the right of data subjects "not to be subject to a decision based solely on automated processing."⁸ Whereas AI can be an assistive tool, it should not be relied on blindly as a crutch. It should be employed in conjunction with the education of users, in a culture where reporting is encouraged, rather than shamed or dismissed, and ultimately as part of a comprehensive approach that purposefully fills the normative void in a principled manner.

7 Whether they are editors, banks (in the case of cryptocurrencies), or institutions more generally in a peer-to-peer environment.

8 Article 22 of the General Data Protection Regulation (GDPR) prohibits decisions that have legal effects and are based solely on automated processing. "Although the regulation applies only within the European Union, it is one the most influential pieces of legislation at the moment and its impact on setting global standards should not be underestimated."

B. Enter the Algorithmic Paradox: “Fueling Polarization,”⁹ Emboldening and Magnifying Extremism

Broadly speaking, algorithms deployed by social-media platforms are engineered to “promote content that will maximize user engagement. Posts that tap into negative, primal emotions like anger or fear, studies have found, perform optimally, thereby nudging people into violence.”¹⁰ Accordingly, and paradoxically, social media “elevates super posters,” giving them unprecedented visibility and disproportionate influence on public opinion. Ironically, merely by reason of their extreme or shocking views, this clusters people into radicalized “ideological bubbles” and creates “skewed perspectives of their own community’s social norms.” Tragically, in Germany and Myanmar, and more recently Pittsburgh (to name just a few examples), that influence has effectively translated into physical violence against immigrants and persecuted groups.¹¹

Post–Cambridge Analytica, it no longer appears controversial to call on platforms that curate “personalized” information to accept a measure of responsibility. That said, it does not naturally follow that we must

9 Anne Applebaum, *Regulate Social Media Now: The Future of Democracy Is at Stake*, WASHINGTON POST, February 1, 2019.

10 Tobias Kraemer *et al.*, *The Good, the Bad, and the Ugly: How Emotions Affect Online Customer Engagement Behavior*. On vitriol, see especially Amanda Taub & Max Fisher, *Facebook Fuelled Anti-Refugee Attacks in Germany, New Research Suggests*, NEW YORK TIMES, August 21, 2018. See also Karsten Müller & Carlo Schwartz, *Fanning the Flames of Hate: Social Media and Hate Crime*, SOCIAL SCIENCE RESEARCH NETWORK, May 21, 2018. See also the Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, April 6, 2018, UN Doc. A/HRC/38/35 (hereinafter: *Kaye Report: Content Regulation*).

11 Max Fisher, *Social Media's Re-engineering Effect, From Myanmar to Germany*, NEW YORK TIMES, November 6, 2018.

task these reticent private actors, equipped chiefly with AI, with this gargantuan chore or endow them with quasi-judicial functions. Doing so would inevitably result in ad hoc, “splendidly indeterminate” criteria¹² for both over- and under-suppression of speech. As Yuval N. Harari presciently observes, “We should [...] fear AI because it will probably always obey its human masters,” assimilating and entrenching their prejudices while allowing them to evade responsibility.¹³ The threat of liability (under the current data-protection construction) provides another misbegotten incentive to suppress more speech than necessary, with little oversight. Hence in the current context, a human rights model—rather than a contractual e-commerce model, in which terms and conditions usurp constitutional rights¹⁴—appears to be more useful than one premised on consent.¹⁵

The policy objectives advanced herein are twofold: First, the law should be returned to the conversation, drawing on supple and principled standards and criteria for platforms to exercise discretion and delist content using the least restrictive means. The restraint of expression (when justified) requires democratically defensible directives.¹⁶ An informal request cannot serve as a basis for the exercise of this authority.

¹² See Rosalie Abella, *Without an Independent Judiciary, Israel's Cherished Democracy Will Be at Risk*, JERUSALEM POST, April 19, 2018.

¹³ Yuval N. Harari, *Why Technology Favors Tyranny*, THE ATLANTIC, October 2018.

¹⁴ See *Douez v. Facebook, Inc.*, US Supreme Court Judgments, [2017] 1 SCR 75, June 23, 2017.

¹⁵ See Helen Nissenbaum, keynote address at the conference on *Artificial Intelligence: Ethical and Legal Implications*, Haifa, Israel, Center for Cyber Law & Policy, University of Haifa (CCLP) and the European Hub, 2018 (with the author).

¹⁶ Regarding the need for authority more generally, see H CJ 5100/94, *Public Committee against Torture in Israel v. Israel* (September 9, 1999) 26.

Second, and flowing from this, in recognition of the internet's borderless nature, there needs to be a better structure of the process by which platforms practice transnational delisting (instead of arbitrary or opaque solutions).¹⁷ Oversight of this practice (in-house, hybrid, or embedded, as in the recent initiative by French regulators, followed by de novo/judicial review) is crucial to this process.¹⁸

To this end, the report follows the following roadmap:

Part One briefly reviews the criteria for limiting hate speech in comparative law and international human rights law. We narrow the discussion to an overview of the main aspects of the law in the United States, as well as in Canada and in Germany (the former chosen as a bridge between the continental and common-law traditions; the latter selected in light of the new legal framework, the *Netzwerkdurchsetzungsgesetz*). We then analyze the approach taken at the international level by judicial and non-judicial bodies, with emphasis on the growing jurisprudence of the European Court of Human Rights, the opinions issued by other agencies of the Council of Europe, and the approach taken by the former UN Special Rapporteur who focused on this matter.

Part Two delineates the nature and scope of the obligations imposed on private actors/platforms with respect to unlawful speech and lays the foundation for recommendations in this regard, based on emerging models. Plainly put, absent a robust normative framework to both guide and scrutinize their decisions, these corporate actors become unwitting

17 The meaning of delisting will be further analysed below.

18 A preliminary example (in its infancy) may be New York City's taskforce on AI, focused on diversity. See City of New York, News Release, *Mayor de Blasio Announces First-In-Nation Task Force to Examine Automated Decision Systems Used by the City*, May 16, 2018.

agents of what the “private surveillance apparatus”¹⁹ has found and thus tend to over-censor speech while missing important instances that require redress.

Mindful of the implications of entrusting and/or unfairly burdening private actors with the regulation of online communications,²⁰ our concluding remarks caution against both maladministration and the inadvertent supplanting of democratic institutions by allowing “data controllers” to usurp their functions. For “if Facebook can control millions of votes with the literal press of a button, is that still democracy?”²¹

19 Nicholas Confessore, *The Unlikely Activists Who Took on Silicon Valley—and Won*, NEW YORK TIMES, August 14, 2018.

20 Including but not limited to partnerships between platforms and foreign governments that raise national security concerns, a treatment of which is beyond the scope of this modest endeavor.

21 Michael Brand, *Can Facebook Influence an Election Result*, THE CONVERSATION, September 27, 2016.

PART I

Identifying Hate Speech

A. National Approaches

1. The United States: Keeping the First Amendment relevant in the digital age

At a certain point in the now infamous “Skokie trial,”²² the judge asked counsel of those opposing the march whether he believed that the town’s Holocaust survivors would be so angered by the planned neo-Nazi demonstration that they could be expected to resort to physical attacks on the marchers, as symbols of their indescribable suffering. “Of course not,” replied counsel. “These people are law-abiding citizens who would never resort to violent attacks.” “Well, then,” the judge is said to have replied, “in that case, I’m afraid there is nothing I can do to prevent the neo-Nazis from marching in Skokie.”²³

²² *Collins v. Smith*, United States Court of Appeals, Seventh Circuit, May 22, 1978, 578 F.2d 1197 (7th Cir. 1978). See also a critique of this approach and of the ACLU’s support thereof in RICHARD DELGADO AND JEAN STEFANCIC, *MUST WE DEFEND NAZIS? HATE SPEECH, PORNOGRAPHY, AND THE NEW FIRST AMENDMENT* 149–62 (1997).

²³ See K. Eltis, “[Absolute Freedom of Expression: The Paradox of the Public Private Distinction](#),” 2003 CU Theses E183 2003. There was no “clear and present danger” of *physical* harm (either by inciting supporters to commit violence or by inciting opponents to harm the marchers. Anecdote shared by Steven Goldstein, teaching American law at the Hebrew University of Jerusalem, 2000 (with the author). See also Kent Greenawalt, *FIGHTING WORDS: INDIVIDUALS, COMMUNITIES AND LIBERTIES OF SPEECH* 53 (1995): “Viewed alone, verbal behavior aimed dominantly at humiliation should not be constitutionally protected

This anecdote epitomizes the intriguing contradictions of present-day First Amendment jurisprudence, ripe for revisiting in the digital age. For in contradistinction to the reluctance generally characterizing the regulation of *any* speech where “state action” is implicated,²⁴ in areas deemed “private” (non-state action), American courts have been surprisingly agreeable to circumscribing a *form* of expression deemed “hostile speech,” in order to satisfy the imperatives of equality.²⁵ In this way they

against punishment (the principle of equalization of victims) [...] that principle, which would protect some victims *not likely to respond with physical force*, recognizes the legitimacy of protecting against deep hurt. [...] The hurt in a particular instance may not correlate with the willingness to fight; indeed, words may hurt the defenseless more than those who are able to strike back" (emphasis added). See also Kent Greenawalt, *Insults and Epithets: Are They Protected Speech?* 42 RUTGERS L. REV. 287 (1990).

24 Mayo Moran, *Talking About Hate Speech: A Rhetorical Analysis of American and Canadian Approaches to the Regulation of Hate Speech*, 1994 WIS. L. REV. 1425. See also Mari J. Matsuda, *Public Response to Racist Speech: Considering the Victim's Story*, 87 MICH. L. REV. 2320 (1989).

25 See Meryl Kirchenbaum, *Hostile Environment, Sexual Harassment Law and the First Amendment: Can the Two Peacefully Coexist?* 12 TEX. J. WOMEN & L. 57 (2002), who addresses the conundrum and notes that, despite the apparent contradictions, "[t]he Supreme Court is not likely to overturn almost twenty years of hostile environment sexual harassment jurisprudence due to a conflict with the First Amendment for practical and policy-based reasons. If the Court ever chooses to address this issue, it will probably find that hostile environment sexual harassment law qualifies as a permissible content-based restriction under one of the various exceptions or limitations stated above" (at 67). Further information can be provided on this point, especially but not limited to Kenneth L. Karst, *Equality as a Central Principle in the First Amendment*, 43 U. CHI. L. REV. 20, 21 (1975); he argues that "[t]he principle of equality, when understood to mean equal liberty, is not just a peripheral support for the freedom of expression, but rather part of the 'central meaning of the First Amendment'"; Wedgwood, *supra* note 2, at 325; CASS R. SUNSTEIN, *DEMOCRACY AND THE PROBLEM OF FREE SPEECH* (1993).

are surprisingly similar to their Canadian and European counterparts, as summarized below.

For our purposes, the workplace is the quintessential example of this approach, which allows some suppression of speech to evade First Amendment scrutiny.²⁶ Because such areas are somehow exempt from the First Amendment's traditional application, claims based on countervailing values underpinning democracy—primarily equality—are given broad credence.²⁷

Significantly, in that “private” arena, hostile speech—as distinguished from any other form of expression protected by the First Amendment—can be and has been curtailed in the United States, insofar as it is deemed to restrict the opportunities of women and minorities to advance on the job. For these limited purposes, the campaign for equality necessarily dictated that the *distinct* character of hostile speech, as differentiated from other forms of expression, and its impact on equality and social participation, be recognized.

²⁶ See, *inter alia*, *Black v. City of Auburn*, 857 F. Supp. 1540, 1549–50 (M.D. Ala. 1994), holding that speech creating a hostile work environment is unprotected; *Robinson v. Jacksonville Shipyards Inc.*, 760 F. Supp. 1486, 1534 (M.D. Fla. 1991), dismissing a First Amendment objection to injunctive relief in a Title VII action; *Jew v. University of Iowa*, 749 F. Supp. 946, 961 (S.D. Iowa 1990).

²⁷ *Id.*; and, *inter alia*, *Baty v. Willamette Industries Inc.*, 172 F.3d 1232 (10th Cir. 1999), upholding the district court's decision to impose liability on an employer for a hostile work environment and rejecting the argument that the sexually harassing speech was protected by the First Amendment. *Robinson v. Jacksonville Shipyards Inc.*, 760 F. Supp. 1486 (M.D. Fla. 1991) stands for the same principle. In the academic context, see, e.g., *West v. Derby Unified School District*, No. 260, 206 F.3d 1358 (10th Cir. 2000), which upheld a school's racial harassment policy under the First Amendment.

Acknowledging—albeit implicitly—the unique effect of hostile speech on both individuals and society in turn allowed American courts to engage in what oddly resembles a balancing act, weighing the [alleged offending] speaker’s right against a countervailing value—equality.²⁸ Upon closer examination, going beyond the rhetoric and established dogma, courts in the United States have recognized equality as a value capable of justifying restrictions on hate speech. This balance can provide valuable insights for us in the digital age.

Befitting the prevalent “Me Too” *Zeitgeist*, John A. Powell (who holds the Chair in Equity and Inclusion at the University of California, Berkeley [and spells his name without capitals]), draws telling analogies between sexual harassment and hate in the digital realm.²⁹

In the nineteen-seventies, when women entered the workplace in large numbers, some male bosses made salacious comments, or hung pornographic images on the walls. “These days, we’d say, ‘That’s a hostile workplace, that’s sexual harassment’” Powell said. “But those weren’t recognized legal concepts yet. So the courts’ response was ‘Sorry, nothing we can do. Pornographic posters are speech. If women don’t like it, they can put up their own posters.’³⁰

28 See Robert Austin Ruescher, *Saving Title VII: Using Intent to Distinguish Harassment from Expression*, 43 REV. LITIG. 249 (2004), citing *Hill v. Colorado*, 530 U.S. 703, 708 (2000).

29 As did Canada’s Supreme Court before him in *R v. Keegstra*, 3 SCR 697 (1990) [hereinafter: *Keegstra*].

30 Andrew Marantz, *How Social-Media Trolls Turned U.C. Berkeley into a Free-Speech Circus*, THE NEW YORKER, July 2, 2018.

Today, the law in the United States, as in most democracies, routinely suppresses sexual harassment in an effort to uphold human rights, particularly the right to equality. Yet it has historically dismissed online racism, notwithstanding the fact that hate speech is not only a prelude to harm³¹ but may constitute harm per se.³² Powell added, “the knee-jerk response [in the United States] is ‘Nothing we can do, it’s speech.’ Well, hold on, what about the harm they’re causing? ‘What harm? It’s just words.’ That might sound intuitive to us now.”³³ But this holds true even more so when the internet—once the bastion of democratic expression—is being increasingly used to “strengthen authoritarian states and weaken or even ‘hack’ democracy.”³⁴ Indicative of this shift, Tim Wu and others increasingly question whether the First Amendment (in its current form) may be rendered “obsolete.”³⁵ The same has been said of the blanket immunity provided by the Communications Decency Act of 1996 (CDA 230).³⁶ This emerging skepticism bodes well for the development of

31 See *Keegstra*, *supra* note 29; Gregory H. Stanton, *The Eight Stages of Genocide*.

32 He argues that *Brown v. Board of Education*, 347 US 483, recognized racism as psychological harm.

33 *Id.*

34 Karen Kornbluh, *The Internet's Lost Promise and How America Can Restore It*, FOREIGN AFFAIRS, WORLD WAR WEB: THE FIGHT FOR THE INTERNET'S FUTURE (September/October 2018), at 37.

35 Tim Wu, *Is the First Amendment Obsolete?* Knight First Amendment Institute (September 2017).

36 Responding to discussions following revelations of Russian interference in the 2016 US presidential election and calls for regulation to stop the spread of disinformation, Tim Hwang said that “these interventions will confront the long-standing legal protections provided by Section 230 of the Communications Decency Act of 1996 (CDA 230), a key legal provision which broadly shields platforms from legal liability for the actions of third-party users

a principled standard for the online environment, one that is mindful of both content and culture.

As a result, notwithstanding the general jurisprudential view that a vigorous First Amendment is of the essence, hate speech has come to be recognized as distinct from ordinary “free speech issues”—however narrowly—in the “private” sphere. In consequence, and curiously, racist or sexist speech is deemed worthy of separate treatment only insofar as it restricts economic prospects within a corporation or educational opportunities on campus.³⁷

of their services. For the past two decades, this provision has been seen as major driver in the growth of online services, and a cornerstone supporting free expression on the web. Simultaneously, CDA 230 has also been argued to inhibit platform responsiveness to the harms posed by harassment, defamation, sex trafficking, and a host of other activities online. The present-day debates on how to address 'fake news' will join the legacy of efforts to reform or eliminate the shield provided by CDA 230" (Tim Hwang, *Dealing with Disinformation: Evaluating the Case for CDA 230 Amendment*. See also Olivier Sylvain, *Intermediary Design Duties*, 50 CONN. L. REV. 203 (2018).

³⁷ That is to say, within the bounds of the realm deemed "private," and in the absence of state action. Thus far it appears that the readiness to recognize the distinct nature of hate speech in the workplace or "private realm," for the purpose of promoting equality, has not extended to public forays, but is instead limited to those areas that the First Amendment overlooks. These areas are highlighted by Frederick Schauer, *The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience*, 117 HARV. L. REV. 1765 (2003): "[T]he law of sexual harassment, which, in both its quid pro quo and hostile-environment aspects, regulates speech, but which, with Supreme Court approval and occasional anguish by commentators, remains unencumbered by the First Amendment's constraints)." An explanation for why this "vast universe of widely accepted content-based restrictions on communication" appears immune to traditional constitutional scrutiny, while an absolutist construction of speech thrives elsewhere, can best be described as elusive, in the absence of commentary addressing this apparent paradox.

For our purposes, the increasing recognition of the unique character of hostile speech directed primarily at women and historically powerless groups online,³⁸ similar to an earlier generation's experience in the workplace, mandates a similar awakening. Plainly put, if American courts no longer shy away from recognizing the insidious character of hostile speech (as a separate category of expression), and if its undeniable impact on equality is taken to justify restrictions on the average citizen's "free speech" in the workplace or on college campuses—the modern locus of the lion's share of popular expression³⁹—a fortiori should they not be reluctant to proceed similarly when it comes the digital "town square."

While it may be premature to draw any definitive conclusions, some measure of rethinking in this context may already be afoot.⁴⁰ Leading jurists such as Carol Christ now recognize what was once unthinkable, namely, that "speech is fundamentally different in the digital context." Similarly, Erwin Chemerinsky, the current dean of Berkeley Law, admits that in the digital age "there is no guarantee that the marketplace of ideas will lead to truth, and that's obviously a big problem."⁴¹

The objective of free speech, as Jack Balkin of Yale convincingly asserts, is above all "to protect and foster a democratic culture in which individuals have a fair opportunity to participate in the forms of meaning-making and mutual influence that constitute them as individuals."⁴²

38 See, e.g., Maggie Astor, *For Female Candidates, Harassment and Threats Come Every Day*, NEW YORK TIMES, August 28, 2018.

39 See *infra*, Part I-B.

40 SUNSTEIN, *supra* note 25. See also Alon Harel's review of the book, 74 B.U.L. REV. 687 (1994).

41 *The Free Speech-Hate Speech Trade-Off*, NEW YORK TIMES, September 13, 2017.

42 J. Balkin, *Digital Speech and Democratic Culture: A Theory of*

Therefore, even after the arrival of the internet, Owen Fiss of Yale Law School cautioned that “the speaker on the street corner, the romantic hero of free speech mythology, has been overtaken by media giants like CBS.”⁴³ A fortiori must we in the digital age “embrace (or at least reluctantly tolerate) state intervention for the purpose of promoting equality. [...] We must learn to embrace a truth that is full of irony and contradiction: that the state can be both an enemy and a friend of speech; that it can do terrible things to undermine democracy but some wonderful things to enhance it as well.”⁴⁴

Surely that is a lesser evil than informally outsourcing this precarious task to companies that fear liability, to secret algorithms, or to foreign governments. Indeed, as Kenneth Lasson powerfully argues in a different context, “a rule of absolute construction cannot be justified merely by asserting that it is too difficult to draw a line between acceptable and unacceptable speech, or too dangerous to entrust the state with making any such distinctions. Such facile abdication of a moral responsibility would deny that there are certain ‘natural laws,’ ‘self-evident truths’ and ‘inalienable rights’—neither opinions nor rebuttable presumptions—upon which the nation was founded and the Constitution based.”⁴⁵

Few today would deny that the internet (not unlike the workplace or school campus) has replaced the traditional “public forum” as the outlet par excellence for expression, and is thus expected to comply with the

Freedom of Expression for the Information Society, 79 N.Y.U. L. REV. (2004).

43 *Id.*

44 Paul Horwitz, *Review of Owen Fiss, THE IRONY OF FREE SPEECH AND LIBERALISM DIVIDED*, 43 MCGILL L.J. 445 (1998).

45 Currently at the University of Baltimore. See Kenneth Lasson, *To Stimulate, Provoke, or Incite? Hate Speech and the First Amendment*, III ST. THOMAS LAW FORUM 49 (1991).

rationales of self-fulfillment and personal autonomy that are commonly invoked as justifying the sacred place of freedom of speech among human rights.

As Danielle Citron of Boston University notes, “The current environment of perfect impunity for platforms deliberately facilitating online abuse is not a win for free speech, because harassers speak unhindered while the harassed withdraw from online interactions. [...] Such abuse should be understood for what it is: a civil rights violation. Our civil rights laws and tradition protect an individual’s right to pursue life’s crucial endeavors free from unjust discrimination. Those endeavors include the ability to make a living, to obtain an education, to engage in civic activities, and to express oneself—without the fear of bias-motivated threats, harassment, privacy invasions, and intimidation.”⁴⁶

2. Canada

Democracies such as Germany and Canada have come to recognize that certain forms of speech disproportionately impair equality, dignity, security, and even speech itself by restricting the opportunities of women and minorities to advance in society. This acknowledgement in turn justifies balancing the inciter’s right to free expression against these other values.⁴⁷ Animated by the imperatives of substantive democracy, and cognizant that “the Holocaust did not begin in the gas chambers but

⁴⁶ Danielle Keats Citron and Benjamin Wittes, *The Problem Isn't Just Backstage: Revising Section 230 Immunity*, 2 GEO L. TECH REV. 453 (2018).

⁴⁷ As George P. Fletcher notes, “the principle of equality under law is best grounded in a holistic view of human dignity.” See *In God's Image: The Religious Imperative of Equality Under Law*, 99 COLUMBIA L. REV. 1608 (1999).

with words,"⁴⁸ the Supreme Court of Canada⁴⁹ deems only unreasonable infringements of constitutional rights (including but not limited to expression) to be unconstitutional under the Canadian Charter of Rights and Freedoms.

48 See Michel Troper, *La Loi Gayssot et la Constitution*, 54 ANNALES HISTOIRE, SCIENCES SOCIALES 1239 (1999); see the explanation that the speech right "must [...] give way to other no less precious rights. A liberty [such as speech] quickly finds its limits, either in others' rights or in the necessities of public order."

49 See *Keegstra*, *supra* note 29:

[A]pplying the *Charter* to the legislation challenged in this appeal reveals important differences between Canadian and American constitutional perspectives. I have already discussed in some detail the special role of s. 1 in determining the protective scope of *Charter* rights and freedoms. Section 1 has no equivalent in the United States, a fact previously alluded to by this Court in selectively utilizing American constitutional jurisprudence (See, e.g., *Re B.C. Motor Vehicle Act*, [1985] 2 S.C.R. 486 (Can.), *per* Lamer J., at p. 498). Of course, American experience should never be rejected simply because the *Charter* contains a balancing provision, for it is well known that American courts have fashioned compromises between conflicting interests despite what appears to be the absolute guarantee of constitutional rights. Where s. 1 operates to accentuate a uniquely Canadian vision of a free and democratic society, however, we must not hesitate to depart from the path taken in the United States. Far from requiring a less solicitous protection of *Charter* rights and freedoms, such independence of vision protects these rights and freedoms in a different way. As will be seen below, in my view the international commitment to eradicate hate propaganda and, most importantly, the special role given equality and multiculturalism in the Canadian Constitution necessitate a departure from the view, reasonably prevalent in America at present, that the suppression of hate propaganda is incompatible with the guarantee of free expression.

See also Jamie Cameron, *Language as Violence v. Freedom of Expression: Canadian and American Perspectives on Group Defamation*, 37 *BUFF. L. REV.* 337, 344, 353 (1988–89).

Thus, for instance, in *Keegstra* that court held that although subsection 319(2) of the Canadian Criminal Code,⁵⁰ which sets out the penalties for the willful promotion of hatred, did in fact violate the Charter's freedom of expression guarantee (because hate speech is a protected form of expression),⁵¹ it nonetheless passed constitutional muster as a reasonable limit under section 1 (the justification clause),⁵² given minorities' right to protection against group-vilifying speech, inter alia.⁵³ The peril of hate speech and, perhaps more importantly, its direct role in sparking genocide,

50 Briefly, four offences in the Canadian criminal code deal specifically with hate speech: (1) s. 318, relating to the advocacy of genocide; (2) s. 319(1), involving the public incitement of hatred; (3) s. 319(2), willful promotion of hatred; and (4) s. 181. The so-called spreading false news provision was struck down as unconstitutional by the Supreme Court of Canada, *R. v. Zundel* [1992] 2 S.C.R. 731, 95 D.L.R. (4th) 202; the majority accepted the proposition that spreading false news could have value.

51 See *Keegstra*, *supra* note 29.

52 The *Canadian Charter of Rights and Freedoms* guarantees the rights and freedoms set out in it subject only to such reasonable limits prescribed by law as can be demonstrably justified in a free and democratic society.

53 *Id.*

In my opinion, a response of humiliation and degradation from an individual targeted by hate propaganda is to be expected. A person's sense of human dignity and belonging to the community at large is closely linked to the concern and respect accorded the groups to which he or she belongs (see ISAIAH BERLIN, *Two Concepts of Liberty*, in *FOUR ESSAYS ON LIBERTY* (1969), 118, at 155). The derision, hostility and abuse encouraged by hate propaganda therefore have a severely negative impact on the individual's sense of self-worth and acceptance. This impact may cause target group members to take drastic measures in reaction, perhaps avoiding activities which bring them into contact with non-group members or adopting attitudes and postures directed towards blending in with the majority. Such consequences bear heavily in a nation that prides itself on tolerance and the fostering of

human dignity through, among other things, respect for the many racial, religious and cultural groups in our society.

A second harmful effect of hate propaganda which is of pressing and substantial concern is its influence upon society at large. The Cohen Committee noted that individuals can be persuaded to believe *almost anything* (p. 30) if information or ideas are communicated using the right technique and in the proper circumstances (at p. 8): [...] We are less confident in the 20th century that the critical faculties of individuals will be brought to bear on the speech and writing which is directed at them. In the 18th and 19th centuries, there was a widespread belief that man was a rational creature, and that if his mind was trained and liberated from superstition by education, he would always distinguish truth from falsehood, good from evil. So Milton, who said *let truth and falsehood grapple: who ever knew truth put to the worse in a free and open encounter*.

We cannot share this faith today in such a simple form. While holding that over the long run, the human mind is repelled by blatant falsehood and seeks the good, it is too often true, in the short run, that emotion displaces reason and individuals perversely reject the demonstrations of truth put before them and forsake the good they know. The successes of modern advertising, the triumphs of impudent propaganda such as Hitler's, have qualified sharply our belief in the rationality of man. We know that under strain and pressure in times of irritation and frustration, the individual is swayed and even swept away by hysterical, emotional appeals. We act irresponsibly if we ignore the way in which emotion can drive reason from the field.

It is thus not inconceivable that the active dissemination of hate propaganda can attract individuals to its cause, and in the process create serious discord between various cultural groups in society. Moreover, the alteration of views held by the recipients of hate propaganda may occur subtly, and is not always attendant upon conscious acceptance of the communicated ideas. Even if the message of hate propaganda is outwardly rejected, there is evidence that its premise of racial or religious inferiority may persist in a recipient's mind as an idea that holds some truth, an incipient effect not to be entirely discounted. (J. E. BICKENBACH,

were again recognized by the Canadian Supreme Court in *Mugesera*.⁵⁴ Although the decision was handed down in the context of immigration (Mugesera was an inciter of the Rwandan genocide who sought refuge in Canada), the recognition that, in the Court's opinion, the danger of hate speech "lies not only in the injury to the self-dignity of target group members but also in the credence that may be given to the speech, which may promote discrimination and even violence," is of the essence. More recently and in the same vein, the Federal Court of Canada recognized that "the damage caused by hate messages to the groups targeted is very often difficult to repair and all the more so in the digital age."⁵⁵

In the end, the curtailment of some speech is intended to protect not only dignity and equality, but other speech and indeed life and personal security, serving as a shield and not as a sword.⁵⁶ The purpose of a reasonable curtailment of hate speech and nefarious misinformation is to enhance the participation in society of all individuals and groups. Just as with sexual harassment, allowing hate speech to stand, in the words of the Rt. Hon. Lord Singh, "undermines democracy by propagating ideas anathema to the democratic ideal—that everyone counts and no one counts more; hate speech is an abuse of right that strikes at the heart of democracy."⁵⁷

KEITH C. CULVER, & MICHAEL GIUDICE [EDS.], *CANADIAN CASES IN THE PHILOSOPHY OF LAW* (5th ed., 2018).

54 *Mugesera v. Canada* (Minister of Citizenship and Immigration), [2005] S.C.C. 40 §8. See Joseph Rikhof, *Hate Speech and International Criminal Law*, 3 J. INT'L CRIM. JUST. 1121 (2005).

55 *Canadian Human Rights Commission v. Winnicki 2005*, FC 1493 (at paragraph 30).

56 Unlike sec. 13, which was repealed for that very reason. See *Hate Speech No Longer Part of Canada's Human Rights Act*, NATIONAL POST, June 27, 2013.

57 4^{eme} Conférence annuelle Chevette-Marx, WHAT IS DEMOCRATIC SOCIETY? Faculté de droit, Université de Montréal, September 6, 2018.

3. Germany

In the United States, hate speech in general is protected under the First Amendment. Hence Congress has been hesitant to enact a law against it, despite the broad bipartisan disgust with the spread of hate-tinged content online.

In Europe, by contrast, there is a greater appetite to regulate speech. In January 2018, Germany began enforcing new rules that levy fines on online service providers that fail to take down hate speech within a day. The European Union, meanwhile, has recently floated the idea of imposing steep penalties on sites that fail to spot and take down terrorist content within an hour.

In a certain sense, the EU approach resembles its German counterpart (from which the proportionality analysis is largely borrowed⁵⁸). Simply put, democracies are duty-bound to take corrective action not to only prevent infringement of the freedom of speech of inciters (as most constitutional democracies and their institutions have done already), but also to protect victims' affirmative rights to expression, dignity, equality, and, ultimately, life and security, in order to survive and thrive. This concept (*wehrhafte Demokratie*) may be said to focus not only on competing rights of various sorts but also on restriction of the freedom of speech in order to safeguard the same right for others.

58 See Aharon Barak, *Proportionality*, in THE OXFORD HANDBOOK OF COMPARATIVE CONSTITUTIONAL LAW (Michel Rosenfeld and András Sajó eds., 2012), 738, at 743. "*Reasonableness in la strongl sense strikes a proper balance among the relevant considerations, and it does not differ substantively from proportionality*). See also Geoffrey Conrad, *Le critère de proportionnalité de l'article premier de la Charte Canadienne: regards historiques et critiques* (2019) (on file with author).

The concern, however, seems to be the apparent dissonance between this important principle and the reality of a controversial new German law, the Network Enforcement Act or *Netzwerkdurchsetzungsgesetz*. The NetzDG, as it is known, is controversial because it compels platforms to deindex “obviously illegal” cyber context. The problem, as previously alluded to and as further discussed here in Part II, is that this law places a heavy onus on private platforms to determine what indeed is “obviously illegal,” not only domestically but also transnationally. As noted below, this may constitute an unfortunate and unintended incentive to overreach, over-chill speech, and succumb to complaints as a function of their zeal rather than their objective merit.

As the *Washington Post* recently warned, “the most problematic issue is that the new law tasks private companies, not judges, with the responsibility to decide whether questionable content is in fact unlawful. In other words, the state has privatized one of its key duties: enforcing the law.”⁵⁹

B. Human Rights Approaches to Hate Speech

In one of his reports, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression called for states and companies to apply international human rights law at all stages of online content regulation: from establishing rules about

⁵⁹ Since the law went into effect, social-media platforms with more than two million users in Germany have been required to erase posts that run afoul of German hate-speech laws. If they fail to delete user-generated content containing Nazi symbolism, denials of the Holocaust, incitements to racial hatred – or a plethora of less clearly defined transgressions, including “insult and blasphemy” – within 24 hours, the companies can be fined up to 50 million euros (\$60 million) and their German executives can be personally liable for five million euros.

what content should be taken down to providing remedies for people harmed by moderation decisions.⁶⁰ The human rights framework is meant to ensure that all individuals are equal and do not suffer any form of discrimination. All individuals are entitled to the same enjoyment of all rights, without distinction of any kind, including on the basis of race, color, sex, language, religion, political or other opinions, national or social origin, property, birth, or any other status.⁶¹ Within this framework, it is recognized that the rights of others are undermined when “deep-rooted hatred is manifested and expressed under certain circumstances.”⁶² It follows, therefore, that states bear an obligation to take measures to limit at least some forms of hate speech.

However, which forms of hate speech and when and how it should be limited are not explicitly and clearly stipulated by the human rights law regime. International human rights law is not a unified, consistent, or robust system of rules that apply universally.⁶³ A positive obligation to regulate hate speech is not founded on any specific provision. On the one hand, certain provisions regulate and limit specific forms of hate speech; on the other hand, certain limitations on speech are justified under the right to freedom of expression. Despite the lack of a universal approach, human rights law provides useful guidance and tools for further consideration.

60 *Kaye Report: Content Regulation* (*supra* note 10).

61 Article 2, *Universal Declaration of Human Rights*, GA Res 217 A, December 10, 1948.

62 *Promotion and Protection of the Right to Freedom of Opinion and Expression*, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue, UN Doc. A/67/357, September 7, 2012, para 37 (hereinafter: *La Rue Report*).

63 See Evelyn Douek, *U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for "Human Rights by Default,"* LAWFARE, June 6, 2018.

1. The Basis for Limiting Hate Speech

There are currently specific provisions of human rights law, at the global and regional levels, that directly prohibit certain forms of hate speech. Such provisions are not found in all human rights treaties and, as already implied, they do not necessarily cover all forms of hate speech. In fact, they might not even label it as such. However, they certainly lay the groundwork for recognizing that states have a positive obligation to limit certain forms of hate speech. Among the international human rights instruments, the International Covenant on Civil and Political Rights states that “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.”⁶⁴ Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination declares that all propaganda activities that promote and incite racial discrimination are illegal.⁶⁵

In addition, provisions of specific treaties require the criminalization of certain forms of hate speech, such as incitement to genocide.⁶⁶ But the

⁶⁴ Article 20 of the *International Covenant on Civil and Political Rights* (ICCPR; adopted December 16, 1966; 999 UNTS 171) puts a strong emphasis on incitement to violence. Similarly, at the regional level, the *American Convention on Human Rights* explicitly prohibits advocacy of national, racial, or religious hatred. Article 13(5) prohibits “any propaganda for war and any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence or to any other similar action against any person or group of persons on any grounds including those of race, color, religion, language, or national origin” (*American Convention on Human Rights* [ACHR, the “Pact of San José”], San José, Costa Rica, adopted November 22, 1969).

⁶⁵ *International Convention on the Elimination of All Forms of Racial Discrimination* (adopted March 7, 1966; 660 UNTS 195).

⁶⁶ Article III(c) of the *Convention on the Prevention and Punishment of the Crime of Genocide* (adopted December 9, 1948; 78 UNTS 277)

Genocide Convention is helpful neither for a definition of hate speech, because it covers a very specific form of hate speech—that which incites to violence and has a special genocidal intent⁶⁷—nor as a regulatory model, because the only obligation that it imposes on states is to criminalize such acts. The extremely high threshold inevitably excludes the regulation of forms of online hate speech that do not reach it.⁶⁸ It is an indicator, though, of a specific form of hate speech that is universally deemed unacceptable.

In Europe, the point of departure for dealing with instances of hate speech has been diametrically opposed to that in the United States. Many continental instruments recognize “hate speech” as a form of expression that must be actively limited, irrespective of whether it occurs in a work context, traditional media outlet, or new internet form of communication such as a social platform.⁶⁹ Therefore, unlike in the United States, the medium through which and the context within which a form of expression is identified are not a *sine qua non* for the qualification of an expression as hate speech.

The European Court of Human Rights (“European Court”) has been the monitoring (and judicial) body with the most abundant jurisprudence on hate speech. The European Court has divided instances of hate speech into those that do not deserve the protection of the European Convention on Human Rights (“European Convention”), on the one

provides that direct and public incitement to commit genocide is to be punishable as a criminal offence.

67 Article 6, *Rome Statute of the International Criminal Court* (adopted July 17, 1998; 2187 UNTS 3).

68 Putting aside valid concerns that even the regulation of such forms of hate speech has not been translated into national law.

69 The Council of Europe, *Hate Speech, Freedom of Expression*. See also the European Commission, *The EU Code of Conduct on Countering Illegal Hate Speech Online*, May 2016.

hand, and those that are not apt to destroy the fundamental values of the Convention but should be restricted nevertheless, on the other hand (the setting-restrictions approach). The first category is based on Article 17 of the European Convention, which provides that acts “aimed at the destruction of any of the rights or freedoms” are not protected by it.⁷⁰ This is the exclusionary approach that, according to the European Court, does not require a balancing test between interference with free expression and pursuit of a legitimate aim. Individuals expressing such views may be deprived of their right to freedom of expression.⁷¹ Alternatively, the European Court has justified limitations on hate speech on the basis of paragraph 2 of Article 10 of the European Convention, which protects the right of freedom of expression.

At the global level, the UN Human Rights Committee, the monitoring body of the International Covenant on Civil and Political Rights, relies on Article 19, which protects freedom of opinion and expression, to decide whether any restrictions imposed on hate speech were justified. The International Covenant permits some further restrictions on freedom of speech in Article 20. Article 19 and Article 20 are understood as complementary: Article 20 is considered a form of *lex specialis* in relation to Article 19 and requires a specific response from the state and the prohibition by law of certain forms of hate speech.⁷² Therefore, according to the

⁷⁰ *European Convention for the Protection of Human Rights and Fundamental Freedoms* (ECHR) (as amended by subsequent protocols), November 4, 1950 (CETS No 5, Rome, Council of Europe). Article 17 was included so that individuals or groups would not be able to hijack the Convention. The exclusionary approach is the polar antithesis of the US approach.

⁷¹ In such cases the restriction of freedom of expression does not constitute interference with freedom of expression and therefore there is no violation of Article 10 of the European Convention.

⁷² ICCPR Committee, General Comment 34, Article 19, Freedoms of

Committee, any restriction under Article 20 must be also justified under Article 19(3).⁷³

Permissible restriction of freedom of expression is the legal basis that has been most often invoked to regulate hate speech. It is indeed the basis for a broader interpretation of the forms of expression that constitute hate speech. States have signed various human rights treaties and thereby committed themselves to protecting freedom of expression, which includes the freedom to hold opinions and the freedom to seek, receive, and impart information and ideas of all kinds, regardless of borders and via every medium.⁷⁴ However, freedom of expression is not an absolute right and is subject to limitations: To be legitimate, any such restriction must be provided for by law and be necessary for the respect of the rights or reputation of others, or the protection of national security, public order, public health, or morals.⁷⁵ It is under these headings that states are understood to have an obligation to interfere with freedom of expression and limit hate speech.⁷⁶

Opinion and Expression, September 12, 2011 (CCPR/C/GC/34), paras. 50–52. *See also* Human Rights Committee, General Comment No. 11, adopted July 29, 1983, Article 20, "Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred," para. 2 (about the relationship between articles 19 and 20). *See also* "Towards an Interpretation of Article 20 of the ICCPR: Thresholds for the Prohibition of Incitement to Hatred: Work in Progress," a study prepared for the regional expert meeting on Article 20 organized by the Office of the United Nations High Commissioner for Human Rights, held in Vienna on February 8–9, 2010.

73 *Id.*, para. 50.

74 ICCPR, Article 19; ECHR, Article 10; ACHR, Article 13. *See also La Rue Report*, *supra* note 62, para. 35.

75 ICCPR, Article 19(3); ECHR, Article 10(2); ACHR, Article 13.

76 "Paragraph 3 expressly states that the exercise of the right to freedom of expression carries with it special duties

Any interference with freedom of expression—even with what is deemed to be hate speech—should be proportionate to and necessary for the legitimate aim it claims to protect. It follows, therefore, that the degree of interference will have to be proportionate to the severity of the statement taken to be hate speech.⁷⁷ This approach recognizes that there is not a single form of hate speech, but gradations of it. As the Human Rights Committee underlined, restrictions must be appropriate to achieving their protective functions and employ the least intrusive instrument available for accomplishing this.⁷⁸ In each and every case, states need to find ways to “reconcile the need to protect and promote the right to freedom of opinion and expression, on the one hand, and to combat discrimination and incitement to hatred, on the other.”⁷⁹

Human rights law offers two regulatory solutions to hate speech. The first calls for the direct prohibition of certain forms of hate speech and consequently permits the direct restriction of certain expressions. However, there is no uniformity even with respect to what forms of expression should be restricted. On the one hand, some provisions, such as Article 20 of the International Covenant on Civil and Political Rights, offer a measure of guidance with regard to the types of speech that may be prohibited. On the other hand, it is left entirely to the European

and responsibilities. For this reason two limitative areas of restrictions on the right are permitted, which may relate either to respect of the rights or reputations of others or to the protection of national security or of public order (*ordre public*) or of public health or morals (ICCPR, General Comment 34, para. 21). For additional analysis of the basis of the recent ECHR decision, underlining the prevalence of this reasoning, see, e.g., *ES v. Austria*, Application no. 38450/12, Judgment, Fifth Section, ECtHR, October 25, 2018.

77 ICCPR, General Comment 34, para. 35.

78 *Id.*, para. 34.

79 La Rue report, *supra* note 62, para. 3.

Court to determine the forms of hate speech that fall under Article 17 of the European Convention. The second regulatory solution does not automatically exclude statements that amount to hate speech. Instead, it requires a balancing test and leaves room for the recognition that hate speech is not a uniform concept and that the responses to it should depend on various circumstances.

Is the best solution simply to identify a threshold above which certain forms of hate speech should be automatically removed from social-media platforms, websites, and the like? Or should a balancing test that includes other factors, such as the degree of interference, be favored?⁸⁰ For instance, the European Court has taken into account the severity of the penalty imposed when it assesses whether the interference with freedom of expression was proportionate to the aim pursued.⁸¹ Such a decision is subject to a balancing test that can be measured only on a case-by-case basis.⁸² This evaluation is arguably a process that permits a more careful assessment of each expression, because it provides an additional layer of

80 "[I]nternational law prohibits some forms of speech for their consequences, and not for their content as such" (*id.*, para. 46).

81 *Gündüz v. Turkey*, Application no. 35071/97, Judgment, First Section, ECtHR, December 4, 2003, para. 54; *Jersild v. Denmark*, Application no. 15890/89, Judgment, Grand Chamber, ECtHR, September 23, 1994, para. 35.

82 "Any restriction imposed on the right to freedom of expression [...] must comply with the three-part test of limitations to the right, as stipulated in Article 19(3) of the Covenant. This means that any restriction must be: (a) Provided by law, which is clear, unambiguous, precisely worded and accessible to everyone; (b) Proven by the State as necessary and legitimate to protect the rights or reputation of others; national security or public order, public health or morals; (c) Proven by the State as the least restrictive and proportionate means to achieve the purported aim" (*La Rue Report, supra* note 62, para. 41).

protection beyond the decision as to whether a specific statement is or is not hate speech.

In *Rabbae v. The Netherlands*, the Human Rights Committee did not evaluate whether the statements amounted to hate speech, as it concluded that the applicants had not exhausted domestic remedies. Crucially, though, the monitoring body reiterated that freedom of expression also covers expression that may be regarded as deeply offensive and that any prohibition of free speech must be construed narrowly. It found that there had been no violation of the rights of the three aggrieved Dutch-Moroccans under the International Covenant on Civil and Political Rights, following the 2011 acquittal of Dutch populist Geert Wilders on charges of insulting the Muslim immigrants in the Netherlands and inciting hatred, discrimination, and violence against them.⁸³ The Committee reasoned that the state is not obliged to ensure that a person charged with inciting discrimination, hostility, or violence will invariably be convicted⁸⁴ and found that the Netherlands had taken “necessary and proportionate” measures in order to prohibit Geert’s statements by prosecuting him under its Criminal Code.⁸⁵

When it comes to the regulation of online speech, the first approach appears to be easier to implement, provided that there is some agreement regarding the definition of hate speech. The second approach

⁸³ *Mohamed Rabbae, ABS and NA v. The Netherlands*, UN Doc CCPR/C/117/D/2124/2011, Views, Human Rights Committee, July 14, 2016, para. 10.4.

⁸⁴ *Id.*, para. 10.6. For further discussion, see Annex III, individual opinion (partly concurring and partly dissenting) of committee members Yuval Shany and Sir Nigel Rodley; and Annex IV, individual opinion (concurring) of committee members Sarah Cleveland and Mauro Politi.

⁸⁵ *Id.*, para. 10.7.

is more difficult to implement, because it is more nuanced and requires a balancing test. Either way, though, further clarification is required regarding the identification of statements that amount to hate speech.

2. The Identification of Hate Speech

There is no consensus in the field of human rights law on how states or other entities are expected to evaluate whether a specific expression is hate speech. Mirroring in a way the diversity of approaches at the national level, there has been no unified approach to how instances of hate speech should be tackled and under which circumstances there is a duty to interfere.

a. Attempts at a strict definition of "hate speech"

None of the international human rights treaties or other legally binding documents contains a definition of hate speech. Only the International Covenant on Civil and Political Rights states that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law." Article 20 provides certain useful elements for determining which forms of expression would amount to hate speech. However, this is still quite a broad definition for the regulation of hate speech in the digital realm and requires further clarification based on practice.⁸⁶

⁸⁶ The Human Rights Committee did not examine this in detail in *Rabbae ao v. The Netherlands* (*supra* note 83).

Some of the proposed definitions exist at the European level only and do not seem to be broadly accepted.⁸⁷ Recommendation 97(20) of the Council of Europe's Committee of Ministers declared that "the term 'hate speech' shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."⁸⁸ This arguably serves more as an inspirational statement than a strict definition.

Additionally, the European Commission against Racism and Intolerance (ECRI) has defined hate speech as "any form of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of 'race,' colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status."⁸⁹ Apart from the general lack of agreement with this outside the ECRI, this definition is quite broad and carries the risk of overregulation and normalization of restrictions on freedom of expression, which should still be seen as exceptional.

87 Both definitions have been adopted by organs of the Council of Europe - the first by the political organ and the second by an independent monitoring body.

88 Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech" (adopted by the Committee of Ministers on October 30, 1997 at the 607th meeting of the Ministers' deputies).

89 ECRI General Policy, Recommendation No. 15 on Combating Hate Speech, Council of Europe, December 8, 2015, p. 3.

b. The ambivalent jurisprudence of the European Court of Human Rights and the approach of the Human Rights Committee

The jurisprudence of the European Court of Human Rights is the most abundant of the international human rights bodies. However, the European Court's approach to hate speech is neither easy to grasp nor consistent. First of all, as previously mentioned, the European Court has divided the instances of hate speech between those that do not fit within the ambit of protection by Article 10 of the European Convention on Human Rights (the exclusionary approach) and those that are not liable to destroy the fundamental values of the convention but should still be restricted (the setting-restrictions approach). However, the dividing line between the instances of hate speech that are excluded from the protection of the Convention and those that merit proportionate and necessary restriction is vague. Second, the European Court has not provided a general definition of hate speech. In some of its judgments, it refers to hate speech as including "all forms of expression which spread, incite, promote or justify hatred based on intolerance (including religious intolerance)."⁹⁰ Beyond this broad reference, however, the European Court has opted for a case-by-case evaluation.

The exclusionary approach was initially used to condemn historical negationism, which denotes a specific category of racist comments that both deny the crimes against humanity and genocide committed by the Nazis during the Holocaust and incite to hatred against the Jewish community.⁹¹ Nevertheless, the European Court's jurisprudence

⁹⁰ *Gündüz v. Turkey* (*supra* note 81), para. 40; *Erbakan v. Turkey*, Application no. 59405/00, Judgment, First Section, ECtHR, July 6, 2006, para. 56.

⁹¹ See *Dieudonné M'Bala M'Bala v. France*, Application no. 25239/13, Decision, First Section, ECtHR; *Lehideux and Isorni v. France*,

has evolved to include hate speech that advocates racial and religious discrimination more broadly. The European Court has further excluded from the protection of the European Convention statements that portray the Jews as the source of evil in Russia or that link all Muslims with a grave act of terrorism.⁹²

The Human Rights Committee has taken a different approach to negationism. The monitoring body of the International Covenant on Civil and Political Rights has rejected general bans on expressions of mistaken opinions or incorrect interpretations of past events.⁹³ This, however, did not prevent it from accepting the necessity of certain restrictions of negationist speech on the basis of Article 19(3), which sets the conditions under which limitations to freedom of expression

Application no. 55/1997/839/1045, Judgment, Grand Chamber, ECtHR, September 23, 1998, paras. 47 and 53. The turning point came in *Garaudy v. France*, Application no. 65831/01, Decision, Fourth Section, ECtHR, June 24, 2003. See also *Honsik v. Austria*, Application no. 25062/94, Decision, First Chamber, EComHR, February 27, 1997; *Marais v. France*, Application no. 31159/96, Decision, Plenary, EComHR, June 24, 1996. And, more recently, *Richard Williamson v. Germany*, Application no. 64496/17, Decision, Fifth Section, ECtHR, January 8, 2019.

⁹² *Roj TV A/S v. Denmark*, Application no. 24683/14, Decision, Second Section, ECtHR, April 17, 2018 (incitement to violence and support for terrorist activity); *Belkacem v. Belgium*, Application no. 34367/14, Decision, Second Section, ECtHR, June 27, 2017; *Pavel Ivanov v. Russia*, Application no. 35222/04, Decision, First Section, ECtHR, February 20, 2007 (ethnic hate); *Hans-Jürgen Witzsch v. Germany*, Application no. 7485/03, Decision, First Section, ECtHR, December 13, 2005; *Glimmerveen and Hagenbeek v. The Netherlands*, October 11, 1979 (racial hate).

⁹³ ICCPR, General Comment 34, para. 49. See also Hungary, Concluding Observations of the Human Rights Committee, Consideration of Reports Submitted by State Parties under Article 40 of the Covenant, November 16, 2010 (UN Doc CCPR/C/HUN/CO/5), para. 19; *Maria Vassilari v. Greece*, Communication no. 1570, Views, Human Rights Committee, March 19, 2009.

are permitted. In *Faurisson v. France*, the applicant complained that his conviction for contesting the existence of gas chambers for extermination in Auschwitz constituted a violation of his freedom of expression. The Committee concluded that the restriction on his freedom of speech was legitimate in order to allow the Jewish community to live in society, free of the fear of antisemitism.⁹⁴ It accepted the government's argument that this restriction was necessary to tackle a subtle form of contemporary antisemitism and that, accordingly, Faurisson's freedom of expression had not been violated.⁹⁵

Nevertheless, the European Court has deemed that certain forms of hate speech deserve evaluation under Article 10. In various cases it has concluded that the interference with freedom of expression was justified. Examples include cases of incitement to hatred, distribution of homophobic leaflets, condoning of terrorism, incitement to ethnic hatred, and incitement to racial discrimination.⁹⁶ However, the European Court is not always consistent in its approach and has often treated pure defamation cases as instances of hate speech.⁹⁷

⁹⁴ *Robert Faurisson v. France*, Communication No 550/1993, Views, Human Rights Committee, November 8, 1996, paras. 9.5–9.6.

⁹⁵ *Id.*, paras. 9.7–10.

⁹⁶ See cases listed at Factsheet: Hate Speech, Press Unit, ECtHR, March 2019.

⁹⁷ It suffices to look at the instances included in its Factsheet on hate speech to fully comprehend the confusion regarding hate speech. It seems that the Court began using the rubric of hate speech to cover various instances that do not necessarily amount to hate speech. However, this approach is not helpful. Freedom of expression can be limited for other reasons and not just in cases of hate speech. Alternative justifications for limitations, such as defamation cases, are considered much less controversial and much more broadly accepted. See *Delfi AS v. Estonia*, Application no. 64569/09, Judgment,

In a seminal judgment, the European Court reached a conclusion that partially contradicts its jurisprudence on negationism. In a series of lectures, the applicant had publicly denied that there had been any genocide of the Armenian people by the Ottoman Empire in 1915 and subsequent years. The Switzerland-Armenia Association lodged a criminal complaint and the applicant was convicted for his statements. The Court concluded that his conviction amounted to a violation of Article 10, because it concluded, on the basis of the facts in front of it, that in a democratic society it was not necessary to impose a criminal penalty on the applicant in order to protect the rights of the Armenian community that were at stake in the present case.⁹⁸ For the balancing test, the European Court assessed the nature of the applicant's statements, the context of the interference with freedom of speech (including geographical and historical factors, as well as the time factor), the extent to which the applicant's statements affected the rights of the members of the Armenian community, the existence or lack of consensus among the High Contracting Parties regarding the legislation, whether Switzerland had an obligation to criminalize the denial of genocide, the method employed by the Swiss courts to reach the conclusion, and, finally, the severity of the interference with freedom of speech.⁹⁹

Grand Chamber, ECtHR, June 16, 2015, para. 140; *see also E.S. v. Austria* (*supra*, note 76).

⁹⁸ *Perinçek v. Switzerland*, Application no. 27510/08, Judgment, Grand Chamber, ECtHR, October 15, 2015, paras. 226–282.

⁹⁹ In the particular case, the Court concluded that the applicant's statements related to a matter of public interest and had not amounted to a call for hatred or intolerance, that the context in which they were made had not been marked by heightened tensions or special historical overtones in Switzerland, that the statements could not be regarded as having affected the dignity of the members of the Armenian community to the point of requiring a criminal-law response in Switzerland, that there had been no international-law obligation for Switzerland

Ultimately, it is not clear on what basis the European Court decides to categorize certain statements as so severe that they do not deserve the protection of the European Convention. While recognizing that hate speech statements have different degrees of severity, it is doubtful whether this division is helpful. Perhaps it would be easier to evaluate all statements under Article 10.¹⁰⁰ The severity of hate speech is always taken into account in balance evaluation. In addition, judging instances that fall under other categories already regulated (such as defamation) under the rubric of hate speech is not helpful for the consolidation of the notion of hate speech.

c. Indicators of hate speech

Somewhere between the attempts to produce a strict definition and the inconsistent approach of the European Court, an intermediate approach would be to identify a series of indicators that would facilitate the determination of whether or not a given statement amounts to hate speech. Despite the lack of consistency in practice, there are certain recurring indicators in international human rights law that may be harnessed to guide any process that intends to limit hate speech in the digital age. The evaluation would remain highly contextual, with a case-by-case determination of the character of each statement; but there

to criminalize such statements, that the Swiss courts appeared to have censured the applicant for voicing an opinion that diverged from the established ones in Switzerland, and that the interference had taken the serious form of a criminal conviction (*id.*, para. 280).

¹⁰⁰ See *also* the recommendation by the Special Rapporteur on Freedom of Expression suggesting that the legal framework for regulating hate speech be based by default on human rights standards. He refers to Article 19(3) of the International Covenant on Civil and Political Rights and calls for the principles of legality, necessity and proportionality, as well as non-discrimination, to inform and guide any regulatory framework. See *Kaye Report: Content Regulation* (*supra* note 10), paras. 7 and 44–48.

would also be a kind of checklist to guide such decisions—a set of variable elements that can be combined in each particular case. Restoring context, as noted, is particularly relevant in the digital realm, which by nature tends to decontextualize.

Indeed, the Committee on the Elimination of All Forms of Racial Discrimination has proposed a checklist of several contextual factors: the content and form of speech; the prevalent economic, social, and political climate at the time the speech was made; the speaker’s position or status; the reach of the speech; and finally, its objectives.¹⁰¹ A similar, six-part threshold test was proposed for expressions to be deemed criminal offences during the discussions that produced the Rabat Plan of Action. It was suggested that the context, speaker, intent, content and form, and “extent of the speech act” be taken into account, as well as the likelihood that the speech would incite actual action against the target group, and with what imminence.¹⁰²

These factors underline the contextual nature of any such evaluation, which cannot and should not be ignored when constructing arguments in favor of automated solutions to hate speech online. These lists of indicators need further consideration and could be amended or expanded. Similar indicators, explicit or implicit, can be found in different cases in practice. The following indicators could be used to determine whether

101 While the Committee proposes these factors as a checklist for the designation of forms of conduct as criminal offences, they can be useful in a broader context. For further details on how each of these factors is understood, *see* Committee on the Elimination of Racial Discrimination, General Recommendation No. 35, Combating Racist Hate Speech, September 26, 2013 (CERD/C/GC/35), para. 15.

102 *See* Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial, or religious hatred, January 11, 2013, UN Doc. A/HRC/22/17/Add.4, para. 29.

specific statements amount to hate speech and accordingly require some form of interference.

First, *the nature of the statement* is a crucial element. The words used are inevitably evaluated in the process, although they are not always helpful. Hate speech does not always manifest itself through expressions of hostility or emotion. It will more often be hidden behind the secondary meaning of words, rather than by the direct "I hate you." "Hate speech can be concealed in statements which at a first glance may seem to be rational or normal."¹⁰³ The *overall content and tone of speech* should also be taken into account.

Second, the *speaker's position or status* is a recurring factor that should be taken into account in various instances. Depending on the circumstances, whether the individual is a member of the political opposition group in a specific country or a journalist could have a significant impact in the determination of whether or not an expression is hate speech.¹⁰⁴ In addition, the *speaker's intention* to incite discrimination, hostility, or violence can be taken into account, even though it is more difficult to prove.¹⁰⁵

Other external indicators may be used to prove the intention of the speaker or as autonomous factors. The *status of the persons targeted* by

103 ANNE WEBER, *MANUAL ON HATE SPEECH* 5 (2009).

104 See *Incal v. Turkey*, Application no. 41/1997/825/1031, Judgment, ECtHR, June 9, 1998, para. 46; *Jersild v. Denmark*, para. 31; *Sürek v. Turkey*, Application no. 26682/95, Judgment, ECtHR, July 8, 1999, para. 63. See also *Halis Doğan v. Turkey* (No. 3), Application no. 4119/02, ECtHR, October 10, 2006, para. 36.

105 *La Rue Report* (*supra* note 62), para. 46. The European Court often will ask whether the applicant intended to disseminate racist ideas and opinions through the use of "hate speech" or whether she or he was trying to inform the public on a matter of public interest.

remarks—whether they belong to a minority or marginalized group that is already subject to discrimination—is an important factor. Any contextual assessment may also consider the existence of *patterns of tension between religious or ethnic communities and discrimination against the targeted group*.

A further distinction is that between *historical facts that can be demonstrated* and *value judgments* that cannot be supported by factual elements. In *Garaudy*, the European Court highlighted that “there can be no doubt that denying the reality of clearly established historical facts, such as the Holocaust, as the applicant does in his book, does not constitute historical research akin to a quest for the truth.”¹⁰⁶

Finally, *how and through which means* the hateful statements have been disseminated can be used to evaluate the impact of the remarks and the need to take action to suppress them. For example, a statement released by an individual to a small and restricted group of Facebook users does not carry the same weight as one published on a mainstream website.¹⁰⁷

This is only a brief overview of suggested indicators. The list of course requires further elaboration and robust reflection. However, it constitutes a first step towards the elaboration of guidelines that would support the constructive evaluation of hate-speech statements in both the online and offline worlds.

C. Conclusion: Towards a Contextual Human Rights Approach to the Regulation of Hate Speech?

It bears repeating that the requisite balancing is not between freedom of speech and some other ill-defined interest. It is instead a question

106 *Garaudy v. France* (*supra* note 91).

107 *La Rue Report* (*supra* note 62), para. 46.

of rights versus rights, as well as the equipoise to be achieved between freedom from improper infringements of expression, on the one hand, and the right of the vulnerable to express themselves, security, and equality, on the other: all are integral to substantive democracy. As Canadian law professor Jean-François Gaudreault-DesBiens argues in a different context: “The dilemma [of inhibiting speech] becomes a duty to regulate against abusive forms of expression, because a constitutional democracy cannot tolerate radical denials of the humanity of some of its citizens.”¹⁰⁸

Beyond the near consensus in human rights law that at least certain forms of hate speech should be limited, this brief overview of human rights law approaches—multiple approaches and not just one—found several recurring indicators that can be harnessed to guide the decision as to whether particular digital statements should be characterized as hate speech. In the digital world, potentially the most challenging concern here is evaluating the context of each statement. Duration and different cultures in the borderless realm bestow different meanings. It is not improbable that a statement may be assessed as innocuous at a certain moment in history, but subsequent events and interpretations give it a different meaning and power.¹⁰⁹ Words do not disappear in the digital

108 Jean-Francois Gaudreault-DesBiens, *From Sisyphus's Dilemma to Sisyphus's Duty? A Meditation on the Regulation of Hate Propaganda in Relation to Hate Crimes and Genocide*, MCGILL LAW JOURNAL (2001).

109 The perils of hijacking human rights narratives in the interest of racist incitement are not unprecedented. The lessons of France's Vichy regime – which, as Richard Weisberg demonstrated, appropriated legal language associated with profound pre-existing social values in order to seamlessly subvert those very principles and lay the foundation for their destruction – are informative beyond speech, with an eye towards democratic government. See RICHARD H. WEISBERG, *VICHY LAW AND THE HOLOCAUST IN FRANCE* 12 (1996) (articulating France's challenge in balancing the push for constitutional reform centered around human rights with maintaining political tradition).

realm; they can be traced, resurface, and be recycled, depending not only on the author's intentions but also on the users' objectives.

The digital realm is full of challenges and contradictions. On the one hand, decontextualized statements easily acquire different meanings that are not subject to geographical or chronological limitations. But recent events, such as the repugnant violence against the Rohingya Muslims in Myanmar, demonstrate the importance of both context and culture, in the more traditional sense. Words may travel around the globe, but they can also strike within specific borders directly at those they target. They may appear harmless to algorithms programmed in the West, but hateful and offensive in the culture that produced them. It is difficult to identify the context that is relevant for assessing, first, whether a statement amounts to hate speech, and second, whether it should be interfered with; this difficulty demonstrates the futility of automated solutions to control hate speech. It further challenges the recent trend of pressuring private companies to regulate hate speech themselves, with no judicial or other direction.¹¹⁰ It is not a task they can or should undertake alone or that they can simplistically resolve by implementing rushed or *ad hoc* solutions that do not wield the appropriate tools for tackling issues of such complexity and breadth.¹¹¹ Abundant resources and possibly multi-tiered

¹¹⁰ Nor can the task of safeguarding constitutional rights or fundamental human rights be outsourced, be it in the traditional sense, to private platforms, or to algorithms, as is increasingly the case. Intermediaries enjoy enormous freedom at the moment to decide whether and how to shape online expression. See Danielle Keats Citron, *HATE CRIMES IN CYBERSPACE*, 1453 (2014). See also the idea that a principled approach predicated on proportionality is preferable: Adam Liptak, *Justices Seem Ready to Boost Protection of Digital Privacy*, *NEW YORK TIMES*, November 30, 2017; *R v. Rogers Communications*, 2016 ONSC 70.

¹¹¹ For as US scholar David Cole observes in a different context, "the problem with [an ad hoc] approach is that it does away with the animating idea of the Constitution - namely that it represents a collective commitment to principles. [...] Constitutional theory [...]"

solutions are required to ensure sufficient understanding of domestic contexts and culture.¹¹² It is essential to return the rule of law to the digital realm.

PART II

New Directions for Public and Private Accountability

A. Who Should Govern Takedown?

Thus far we have addressed—however summarily—the imperative of curtailing racist speech in the digital age. But as stated above, the broader and more pressing issue is not whether online speech should be limited or results delisted—inasmuch as this already happens (quite regularly)—but how to subject this practice to the rule of law, in terms of both authorization and oversight. Intermediaries' obligation (a term we prefer to liability), as Daphne Keller puts it, “sits at a unique and often troubling intersection of state and private power.”¹¹³

requires an effort, guided by text, precedent and history, to identify the higher principles that guide us as a society.” See Richard Posner, reply by David Cole, *How to Skip the Constitution: An Exchange*, THE NEW YORK REVIEW OF BOOKS, January 11, 2007.

112 This is why, as a *piste de réflexion*, it might be a good idea to have a two-pronged approach, including a committee at the domestic level (akin to New York City's taskforce on AI; *supra*, note 19) in order to properly understand local context/culture.

113 See Daphne Keller, *The Right Tools: Europe's Intermediary Liability Laws and the EU 2016 General Data Protection Regulation*; Daphne Keller, *SESTA and the Teachings of Intermediary Liability*.

The imperative consequently begins, but does not end, with limiting the disproportionate and frighteningly chilling impact, unprecedented in history, that the digital medium grants racist speakers; but it must also avoid and prevent abusive and increasingly arbitrary takedown requests that are not supported by the law.¹¹⁴

Plainly put, the fear is one of “globalizing censorship,”¹¹⁵ as platforms, their outsourced employees, and algorithms scramble to satisfy (or at the very least appease) European regulators under the current data-protection/liability model. That in turn inevitably leads to overreach and ironically imperils free speech by erring on the side of suppression or indiscriminate censorship.

Examples abound. We will limit ourselves to three of them. As the *New York Times* reported in an article illustrating the “inconsistencies and gender bias” of Facebook’s homemade “takedown” policies, the giant banned an image of a woman’s naked back in an ad for a romance novel, while allowing a man’s bare chest to appear in the very same context.¹¹⁶ Death threats against women thrive online, while an innocuous “men suck” post in a doctoral student’s experiment was promptly removed.¹¹⁷ In an especially ironic twist, Eastern European authors condemning antisemitism and citing the racist passages they called out as repugnant found themselves censored for posting antisemitic

114 On the problems of notice and takedown system, see Keller, *The Right Tools*, and Keller, *SESTA*, *supra* note 113.

115 [*“The Cyberlaw Podcast: Globalizing Censorship,” interview with Karen Eltis*](#), Steptoe Cyberlaw [podcast] episode 168.

116 Sapna Maheshwari & Sheera Frenkel, *Facebook Lets Ads Bare a Man's Chest. A Woman's Back Is Another Matter*, *NEW YORK TIMES*, March 3, 2018.

117 Anastasia Berwald, *Insta-Censure: La censure du féminisme en ligne*, LABORATOIRE DE CYBERJUSTICE.

content.¹¹⁸ The censorship was automatic, based on “perceived legal requirement,” and paradoxically motivated by *the number* of complaints lodged against them and the fear of liability, a “criterion” that inter alia flies in the face of the logic of the marketplace. It is a phenomenon that Jack Balkin eloquently labels “collateral censorship,” where platforms err on the side of caution (or removal) of “even fully protected speech to avoid the spectre of liability.”¹¹⁹

B. Beyond Immunity and Beyond the Burden of Normative Ambiguity

In the United States, under Section 230 of the Communications Decency Act and the Digital Millennium Copyright Act (DMCA), platforms have long enjoyed full immunity for the content they host.¹²⁰ Today, however, in addition to privacy bills advancing in both the House and the Senate which challenge that remarkable status, now increasingly serving as a normative black hole, “once-unthinkable support” from platforms themselves and “shifting consumer attitudes are signaling a chance for momentous change.”¹²¹

For instance, Prof. Jonathan Zittrain, a leading thinker in this area, acknowledges that misleading information (including but not limited to

118 Iris Georlette, *On the Ground: An Eyewitness to Antisemitism in Ukraine – Diaspora*, JERUSALEM POST, June 10, 2018.

119 NEW CONTROVERSIES IN INTERMEDIARY LIABILITY LAW: ESSAY COLLECTION (edited by Tiffany Li), Information Society Project, Yale Law School, Spring 2019.

120 The same is true to a certain extent in Europe under the e-commerce directive. For a detailed discussion, exceeding scope of the present paper, see Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech*, Hoover Institution, Aegis Series Paper 1902 (2019).

121 *Id.*

racist speech) is indeed a problem, but nonetheless rejects the liability-based solution embraced by the EU General Data Protection Regulation (GDPR) and urges us to rethink the full immunity that Section 230 grants intermediaries.¹²² In his words,

The platforms were free to structure their moderation and editing of comments as they pleased, without a traditional newspaper's framework[,] in which to undertake editing was to bear responsibility for what was published. If the *New York Times* included a letter to the editor that defamed someone, the *Times* would be vulnerable to a lawsuit. [...] Not so for online content portals that welcome comments from anywhere—including the online version of the *New York Times*.¹²³

The rationale underlying that exemption, Zittrain recognizes, is no longer appropriate for “an infant industry [that] has grown up,”¹²⁴ in turn reflecting a crack in the “immunity armour”¹²⁵ as well as in the “marketplace of ideas” dogma, as previously discussed.¹²⁶

122 Jonathan Zittrain, *Don't Force Google to Forget*, NEW YORK TIMES, May 14, 2014. See also David Streitfeld, *European Court Lets Users Erase Records on Web*, NEW YORK TIMES, May 13, 2014; Jonathan Zittrain, *CDA 230 Then and Now: Does Intermediary Immunity Keep the Rest of Us Healthy?*, November 10, 2017.

123 *Id.*

124 *Id.*

125 Karen Eltis & Pierre Trudel, *Rapport Canadien: Le déréférencement à l'ère numérique - une approche hybride pour faire le pont entre la vision européenne et américaine du droit à l'oubli* (2019).

126 Another prominent US internet scholar, Prof. Jack Balkin, has entertained the idea of imposing “fiduciary obligations.” See Jack M. Balkin, *Information Fiduciaries and the First Amendment* (2016), Faculty Scholarship Series, 5154; Jonathan Zittrain, *Engineering*

C. The Delegation of Speech Regulation to Private Actors

Recent regulatory initiatives, particularly in Europe, raise serious concerns about the protection of freedom of expression in the digital realm. Because governments are unable to police content directly—due to lack of jurisdiction and/or of an appropriate legal framework—they have begun to enact legislation that increases the pressure on companies to monitor and police online content—what is known as “intermediary responsibility.”¹²⁷ Germany was one of the first countries to pass a specific law regulating online hate speech. The new legislation, which was examined in the first part of this study, requires social-media companies to remove illegal content within 24 hours of its being reported to them. A provider that fails to comply with the law risks a fine of millions of euros.¹²⁸

Unfortunately, more and more governments have been adopting this idea. Russia soon copied the German law but left the definition of unlawful content deliberately vague.¹²⁹ The 2016 European Union Code of Conduct on countering illegal hate speech online involves an agreement between

an Election. Digital Gerrymandering Poses a Threat to Democracy, Response 2014, 127 HARV. L. REV. F. 335.

¹²⁷ We purposely use the word “responsibility” rather than “liability,” the fear of which tends to encourage indiscriminately suppressing content in order to satisfy regulators and avoid penalties as noted above.

¹²⁸ Soon after, the Special Rapporteur on Freedom of Expression sent a letter to Germany expressing his concern that the new law is vague, ambiguous, and could result in overregulation of speech in order to avoid fines. Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, June 1, 2017, OL DEU 1/2017.

¹²⁹ Reporters Without Borders, *Russian Bill Is Copy-and-Paste of Germany's Hate Speech Law*, July 19, 2017. See also Jacob Mchangama

the European Union and four major companies to remove content. The agreement refers to “trusted flaggers” and the development of “counter-narratives.”¹³⁰ Such initiatives run “the risk of transforming platforms into carriers of propaganda well beyond established areas of legitimate concern.”¹³¹

Indeed, such laws seek to consolidate legal obligations already stipulated in previous legislation. Hate speech, for instance, is already prohibited by German law. There is no doubt that governments need to collaborate with companies that host content in order to implement any legal framework to combat hate speech. But these new laws simply delegate the entire regulatory process to the private actors. They do not stipulate indicators to determine what content amounts to hate speech; there is no provision for an oversight by a judicial or other authority; and there are no guarantees of due process. In his recent report, the Special Rapporteur on Freedom of Expression urged states to reconsider speech-based restrictions and adopt smart regulation aimed at enabling the public to decide how and whether to be part of online forums.¹³²

The European Court of Human Rights has been gradually developing a similar but more nuanced approach, which unfortunately reinforces this tendency to strengthen intermediaries’ liability to proactively regulate content. The first case in which the European Court pronounced on the

& Joelle Fiss, *Germany's Online Crackdowns Inspire the World's Dictators*, FOREIGN POLICY, November 6, 2019.

130 *The EU Code of Conduct on Countering Illegal Hate Speech Online*, May 2016.

131 *Kaye Report: Content Regulation* (*supra* note 10), para. 21.

132 *Id.*, paras. 19–21. See also the Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, August 29, 2018, UN Doc A/73/348 (on artificial intelligence).

matter was *Delfi AS v. Estonia*.¹³³ The Court found that the rights and interests of others and of society as a whole may entitle states to impose liability on internet news portals (in this case the company provided a platform for user-generated comments on previously published content) if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties.¹³⁴ In *Pihl v. Sweden*, though, the Court rejected as manifestly ill-founded the individual's claim against a small nonprofit association concerning a blog on which someone had anonymously posted a defamatory comment.¹³⁵ The distinction between "active" hosting of user-generated content and "passive" hosting has been challenged and the lines between the two have been blurred.¹³⁶

133 *Delfi AS v. Estonia*, Application no. 64569/09, Judgment, Grand Chamber, ECtHR, June 16, 2015.

134 The Court took various factors into account, including the extreme nature of the comments in question, the fact that they had been posted in reaction to an article published by the applicant company on its professionally managed news portal run on a commercial basis, the insufficiency of the measures taken by the applicant company to remove comments without delay after publication and the moderate sanction (320 euro) imposed on the applicant company, to conclude that the Estonian court's finding of liability against the applicant company did not amount to a violation of the platform's freedom of expression (*id.*).

135 The Court took into account the fact that the comment, although offensive, had not amounted to hate speech or an incitement to violence; it had been posted on a small blog run by a non-profit association; it had been taken down the day after the applicant had made a complaint; and it had been on the blog for only around nine days (*Pihl v. Sweden*, Application no. 4742/14, ECtHR, March 9, 2017). See also *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary*, Application no. 22947/13, ECtHR, February 2, 2016.

136 For a brief commentary on the distinction, see Aleksandra Kuczerawy, *Active vs. Passive Hosting in the EU Intermediary Liability Regime: Time for a Change?* August 7, 2018.

While the aforementioned court decisions constitute some first steps towards the regulation of one of the most difficult problems in the digital realm, for now they risk increasing the tendency to overregulation and over-censorship, and threaten democratic values and fundamental principles of human rights.

D. Nuanced Approaches to Responsibility

While it is far beyond the scope of this report to dwell on individual legal systems in any detail, the following point, which relates to the increasing convergence of approaches in the common-law world (outside the United States), is worth making when we entertain new directions for intermediary responsibility.

Especially noteworthy is an Australian decision that adopts the reasoning (obiter) of a ruling by the Supreme Court of Canada. *Google Inc. v. Duffy* rejects the “merely hosting” rationale traditionally applied to platforms and takes the position that the operator of a search engine can be held responsible for failing to delist defamatory content when said operator does more than simply providing hyperlinks (e.g., by showing excerpts of the content or via Google’s “autocomplete” function).¹³⁷ This concept, directly inspired by emerging Canadian law, and significant for its cross-border viability, is far more consistent with the rationale underlying delisting in the digital age and achieves a greater balance between freedom of expression and the right to privacy and reputation—the values on which the so-called right to be forgotten is based.¹³⁸ The most recent decision by the European Court of Human Rights, which

¹³⁷ *Google Inc. v. Duffy*, 2017 SASCFC 130 (October 4, 2017).

¹³⁸ GDPR, Article 17; *Google Spain v. AEPD and Mario Costeja González*, C131/12, Court of Justice of the European Union (CJEU), May 13, 2014. See Karen Eltis, *The Anglo-American/Continental Privacy Divide? How Civilian Personality Rights Can Help Reconceptualize*

concluded that a hyperlink is not ipso facto defamatory, should be further analyzed.¹³⁹ That said, suffice it to note that at this juncture the European Court decision may be said to differ from *Duffy* in that it seems to reject the automatic attribution of liability for hyperlinking, unless the context justifies such liability, as the Australian decision found.

Most importantly, this contextual conception, *purposely anchored in substantive human rights rather than in data protection or procedural notions*, makes it possible for politically independent courts, rather than reticent or inexperienced corporate actors (and their AI), to determine what truly constitutes racist incitement or defamation and must therefore be delisted.¹⁴⁰ It stands to reason that this contextual test,¹⁴¹ tapping into time-honored principles of human rights law that go beyond the more procedural (and at times nebulous) data protection model,¹⁴² lends itself far better to the online environment and should be given further consideration.

the "Right to Be Forgotten" towards Greater Transnational Interoperability, 94(2) CANADIAN BAR REVIEW 355 (2016).

139 *Magyar Jeti Zrt v. Hungary*, Application no. 11257/16, Judgment, Fourth Section, ECtHR, December 4, 2018.

140 *C.L. c. BCF Avocats d'affaires*, Commission d'accès à l'information du Québec, 2016 QCCA 114, on the right to rectification under the Québec Act Respecting the Protection of Personal Information in the Private Sector: "The company must take all reasonable means to rectify the plaintiff's information internally (on its internet site), which is not, however, equivalent to a duty to delist (externally, on the rest of the Web)" [translation by authors]. The decision does not address intermediaries' duties.

141 *Duffy v. Google: Is This the End of the Internet as We Know It?* DEFAMATION E-BULLETIN, October 30, 2015.

142 Keller, *supra* note 120.

The debate can be enriched by initiatives such as the 2014 Manila Principles on Intermediary Liability.¹⁴³ Developed by a coalition of civil society experts, they identify key principles that can inform and guide the development of any intermediary liability framework.¹⁴⁴ Due regard to the reports by the Special Rapporteur on Freedom of Expression is also required.¹⁴⁵

E. Concluding Remarks: Convergence of Legal Frameworks; Protecting Freedom of Expression through the Rule of Law

Expression is contextual and cultural. It is above all human. The digital realm and its algorithms, however unintentionally but as a function of the current economic model, decontextualize speech and magnify radicalism. In order to avoid overreaching while at the same time allowing hateful incitement to thrive, the application of standards must—like all regulation—provide the correct incentives. Hence responsibility may prove a better vehicle than the liability-based model. What must we do? The objective, as previously noted, is to preserve the relevance of the law and erect more appropriate boundaries, anchored in legal principles and authority, rather than chilling speech because of informal requests or corporate fears of liability. A binding normative instrument is far preferable in this regard.

143 Manila Principles on Intermediary Liability.

144 *Kaye Report: Content Regulation* (*supra* note 10), para. 14.

145 Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, May 11, 2016, UN Doc A/HRC/32/38 (on intermediary responsibility). *See also* Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, August 29, 2018, UN Doc A/73/348 (on artificial intelligence).

Here the German concept of *Drittwirkung*¹⁴⁶ is helpful. Human rights may be upheld not only against the state but “against any group in society that is sufficiently powerful to functionally substitute for the state.” In the end, therefore, “the state may have an affirmative constitutional responsibility to create private law to protect a citizen against the actions of private groups or individuals,”¹⁴⁷ including in a transnationally viable fashion.

An interesting model that can be built upon is the human rights model.¹⁴⁸ Specifically, like the Canadian “cooperative” ombudsman model (premised on human rights principles rather than on data protection), the nascent French multi-stakeholder model works with platforms towards the proportional limit of hate speech, “embedding” regulators on site (as is done, for instance, in banking and the nuclear industry). While the precise extent of oversight and independence in this case is not yet known, a two-step process (which could be a hybrid process with both internal

146 The German doctrine of Third Party Effect of Fundamental Rights. See generally Eric Allen Engle, *Third Party Effect of Fundamental Rights (Drittwirkung)* (October 1, 2009). 5 HANSE LAW REVIEW 165 (2009). See also DONALD P. KOMMERS & RUSSELL A. MILLER, *THE CONSTITUTIONAL JURISPRUDENCE OF THE FEDERAL REPUBLIC OF GERMANY*, 432 (3rd. ed., 2012).

147 Stephen Gardbaum, *The "Horizontal Effect" of Constitutional Rights*, 102 MICH. L. REV. 387 (2003). Gardbaum writes, inter alia: "Among the most fundamental issues in constitutional law is the scope of application of individual rights provisions and, in particular, their reach into the private sphere. This issue is also currently one of the most important and hotly debated in comparative constitutional law, where it is known under the rubric of 'vertical' and 'horizontal effect.' These alternatives refer to whether constitutional rights regulate only the conduct of governmental actors in their dealings with private individuals (vertical) or also relations between private individuals (horizontal). In recent years, the horizontal position has been adopted to varying degrees, and after systematic scholarly and judicial debate, in Ireland, Canada, Germany, South Africa, and the European Union, among others."

148 See *supra*, note 14.

and external controls) might point to new directions for meaningful collaborative oversight (as in the first Industrial Revolution), leading to “les normes et la Coutume” of the Civil Code.

Going a step beyond the separation of powers issue, “where platforms both create and apply the rules as de facto adjudicators” without the obligation to provide reasons and absent an appointment process, we might apply Raz’s indicia. In his most recent article,¹⁴⁹ Raz reiterates that “at least one [...] aim of the [rule of law] is to avoid arbitrary government.” “Stability and predictability” are essential, as is “observing due process.” Although not set out in the cyber context, the criteria he highlights may lend themselves to fostering greater transparency, reasoning and oversight in the digital realm and making the process verifiable. The indicia must be: “(1) reasonably clear, (2) reasonably stable, (3) publicly available, (4) general rules and standards, that are (5) applied prospectively and not retroactively.”¹⁵⁰

In sum, and as we hope to elaborate in the future at a different stage in our common reflection, the opacity of the deployment of automated or semi-automated decision-making processes to regulate online content calls into question these processes’ legality and legitimacy (and may indeed violate the GDPR’s enshrined right to human control). These concerns must be robustly addressed, perhaps by a transversal¹⁵¹ and cooperative (three-)step human oversight process at various levels, which takes culture and context into account. “To counter not only the spread of high-tech repression abroad but also potential abuses at home, policy makers in democratic states must think seriously about how to mitigate

149 Joseph Raz, *The Law’s Own Virtue*, 39 OXFORD JOURNAL OF LEGAL STUDIES 1 (2019).

150 *Id.*

151 Perhaps taking the form of a cross-border Alternative Dispute Resolution (ADR) mechanism.

harm and shape better practices"¹⁵² and allow historical mechanisms of democratic norms to continue to be applied.

Recommendations

As previously noted, the regulation of online hate speech is particularly complex, because of its context and cultural dependence as well as the borderless nature of the digital realm and its tendency to inadvertently give greater prominence to extreme views. The discussions have focused primarily on pushing the platforms that post the content to remove it as soon as it is flagged or identified. This paper offered a comparative analysis of national approaches to regulating hate speech as well as the human rights approach to hate speech and offered the following recommendations:

1. The adoption of universal “hate speech” indicators informed by human rights law practices

The report highlights a series of indicators for identifying instances of hate speech, gleaned from the existing human rights law regime, which can be assembled into purposive strategy for addressing online hate speech. These indicators are important for consistency, given that, in accordance with the principles of democratic legitimacy, private actors cannot and should not shoulder the burden of “moderating” speech alone or in an *ad hoc* manner.

152 Steven Feldstein, *The Road to Digital Unfreedom: How AI is Reshaping Repression*, 30(1) JOURNAL OF DEMOCRACY 40 (2019).

2. Interim measures: The consideration of less-intrusive measures by platforms and other relevant actors

In practice, certain means of addressing online hate are—by virtue of the very medium—“faster than the speed of government.”¹⁵³ Therefore, platforms and other actors should engage in hate-speech moderation with an eye towards *proactively curtailing* incitement liable to cause irreparable harm (including but not limited to genocide, which, as we know, begins with the rhetoric of hate). Proactive measures will make it possible to avoid post-factum reactions to events and uncontrolled over-censorship. Such interim measures may include the following:

- Harnessing algorithms to curtail the virality of hateful content; in other words, to make online racism less accessible instead of giving it exaggerated prominence;
- Simplifying and making regulation more transparent and visible. For example, platforms could be required to introduce “pop-up” educational videos on online hate speech.

3. Introduce tiers of oversight by distinguishing between “easy” and “hard” cases

Imposing intermediary obligations, as Daphne Keller puts it, “sits at a unique and often troubling intersection of state and private power.”¹⁵⁴ The imperative consequently begins, but does not end, with limiting the disproportionate and frighteningly chilling impact, unprecedented

¹⁵³ Alain Dutoit, “Trust in the Age of Digital Disruption,” Fourth Annual Digital Government Forum, Ottawa, ON, June 20, 2018.

¹⁵⁴ See Keller, *The Right Tools* (*supra*, note 113); Keller, *SESTA and the Teachings of Intermediary Liability* (*supra*, note 113). See also Keller, *supra* note 120.

in history, that the digital medium grants racist speakers, but must also avoid and prevent abusive and increasingly arbitrary takedown requests that are not supported by the law. Oversight must be available at various levels: at the platform level (initial appeal of takedown); a hybrid alternative dispute resolution (ADR) process; and oversight by courts to uphold the rule of law.

- “Easy cases”¹⁵⁵ may be dealt with by platforms more quickly (“think fast” [Daniel Kahneman]), subject to their furnishing their reasons and allowing the possibility of appeal.
- “Hard cases,”¹⁵⁶ which require decision-makers to “think slow,” are best left to an accountable judiciary. This is in line with the GDPR, which requires that certain decisions not be taken without human oversight.¹⁵⁷

4. Ensure the right to appeal any takedown decision, whether taken by platforms or judicial authorities

Human rights law provides a right to receive a remedy for any violation of human rights. Accordingly and as noted, individuals must have the right to receive a reasoned decision for the removal of their statements, as well as the option to appeal such decisions.

155 These, in accordance with Aharon Barak's definition in a more general context, require decision-makers/judges to use little or no discretion, because the rule is clear and “easy” to apply.

156 For Barak, cases that, given their thorniness, require decision-makers/judges to deploy significant discretion.

157 “The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.” GDPR, Article 22(1) – but this exceeds the scope of our limited discussion.

Proposals for Improved Regulation of Harmful Online Content

Susan Benesch¹

Introduction | PART I: Substantive Standards | PART II: Procedural Standards | Appendix

Introduction

In its early life, the internet inspired optimism that it would improve the world and its people, but that has been supplanted by alarm about harmful content, often viral words and images. Though the vast majority

¹ I am very grateful for comments on an earlier version of this paper and its ideas by colleagues in academia, at tech companies, and at NGOs, including Chinmayi Arun, Dan Bateyko, Liz Carolan, Connie Chung, Pierre François Docquir, Rob Faris, Tonei Glavinic, David Kaye, Michael Lwin, Colin Maclay, K. S. Park, and Kit Walsh. Tonei Glavinic also contributed invaluable research, ideas, and editing. Your questions, comments, and critiques are also extremely welcome: sbenesch@cyber.law.harvard.edu

of online content is still innocuous or beneficial, the internet is also polluted by hatred: some individuals and groups suffer harassment or attacks,² while others are exposed to content that inspires them to hate or fear other people, or even to commit mass murder.³

Hateful and harmful messages are so widespread online that the problem is not specific to any culture or country, nor can such content be easily classified under terms like “hate speech” or “extremism”: it is too varied. Even the people who produce harmful content, and their motivations for doing so, are diverse. Online service providers (OSPs)⁴ have built systems to

2 See, e.g., European Commission, Directorate-General for Communication, [Special Eurobarometer 452, Media Pluralism and Democracy](#) (November 2016), at 17, reporting: “A large majority of those who follow or participate in debates has heard, read, seen or themselves experienced cases where abuse, hate speech or threats are directed at journalists/bloggers/people active on social media (75%)”. See also National Society for the Prevention of Cruelty to Children, [Online Abuse: Facts and Statistics](#); Maeve Duggan, [Online Harassment 2017](#), Pew Research Center, July 11, 2017 (reporting a survey in which 62% of US respondents regarded online harassment as a major problem and 40% had experienced it themselves); Steve Stecklow, [Why Facebook Is Losing the War on Hate Speech in Myanmar](#), Reuters, August 15, 2018; United Nations, Human Rights Council, [Detailed Findings of the Independent International Fact-Finding Mission on Myanmar](#), A/HRC/42/CRP.5, September 16, 2019.

3 Jacob Asland Ravndal, [The Online Life of a Modern Terrorist: Anders Behring Breivik's Use of the Internet](#), VOX PoL (October 24, 2014); Jessica Schulberg, Luke O'Brien, & Oliver Mushtare, [The Neo-Nazi Podcaster Next Door](#), HUFFPOST (February 7, 2019); Adam Taylor, [New Zealand Suspect Allegedly Claimed "Brief Contact" with Norwegian Mass Murderer Anders Breivik](#), WASHINGTON POST, March 15, 2019.

4 In this paper, “online service providers” (OSPs) refers to companies that host and disseminate user-generated content online and attempt to limit harmful content. Google, Facebook, and Twitter are the best known and most discussed, but there are many others, large and small, including Reddit, Automattic, Bytedance, and companies

diminish harmful content, but those are inadequate for the complex task at hand and have fundamental flaws that cannot be solved by tweaking the rules, as the companies have been doing so far. The stakeholders who have the least say in how speech is regulated are precisely those who are subject to that regulation: internet users.⁵ “I’ve come to believe that we shouldn’t make so many important decisions about speech on our own,” Mark Zuckerberg, a founder and the CEO of Facebook, wrote last year.⁶ He is correct.

Daunting though the problem is, there are many opportunities for improvement, but they have been largely overlooked. The widespread distress about it is itself an opportunity, since that means millions of people are paying attention, and it will take broad participation to build online norms against harmful content. Such mass participation is neither far-fetched nor unfamiliar: many beneficial campaigns and social movements have been born and developed thanks to mass participation online.⁷

that build and maintain chat apps, niche social-media platforms, or online games.

5 Rebecca MacKinnon developed this idea in a 2012 book, and others have since joined her in calling for some kind of oversight of companies' governance of the speech of billions. *See, e.g.,* REBECCA MACKINNON, *THE CONSENT OF THE NETWORKED* (2012); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, *HARVARD LAW REVIEW* 131 (2017); TARLETON GILLESPIE, *CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA* (2018).

6 Mark Zuckerberg, *The Internet Needs New Rules. Let's Start in These Four Areas*, *WASHINGTON POST*, March 30, 2019.

7 Zeynep Tufekci describes many of these in a 2018 book, though she also points out that the relative ease and speed of mass organizing online can make it harder to sustain social movements. *See* ZEYNEP TUFECKI, *TWITTER AND TEAR GAS, THE POWER AND FRAGILITY OF NETWORKED PROTEST* (2018).

This paper offers a set of specific proposals for better describing harmful content online and for reducing the damage it causes, while protecting freedom of expression. The ideas are mainly meant for OSPs since they regulate the vast majority of online content; taken together they operate the largest system of censorship the world has ever known, controlling more human communication than any government.⁸ Governments, for their part, have tried to berate or force the companies into changing their policies, with limited and often repressive results.⁹ For these reasons, this paper focuses on what OSPs should do to diminish harmful content online.

The proposals focus on the rules that form the basis of each regulation system,¹⁰ as well as on other crucial steps in the regulatory process, such as communicating rules to platform users, giving multiple stakeholders a role in regulation, and enforcing the rules.

8 The only government whose censorship system could rival the companies' in number of users or volume of content regulated is that of China, which has fewer than one billion people online; *see, e.g.*, Jon Russell, *China Reaches 800 Million Internet Users*, TECHCRUNCH (October 21, 2018). Facebook alone has more than 2.3 billion regular monthly users. *see* Facebook. YouTube has nearly two billion users and sees more than 400 hours of video posted every minute; *see* Danielle Abril, *YouTube Nears Major Milestone amid Emphasis on Subscriptions*, FORTUNE, February 4, 2019; Google Inc., *Monetization Systems or "The Algorithm" Explained*, YOUTUBE HELP.

9 WILLIAM ECHIKSON & OLIVIA KNOTT, *GERMANY'S NETZDG: A KEY TEST FOR COMBATING ONLINE HATE* (2018); Anthony Cuthbertson, *Pakistan Lifts Three-Year YouTube Ban on the Condition Censors Can Request Content Removal*, NEWSWEEK, January 19, 2016.

10 I use the terms "moderate" and "regulate" to refer to the OSPs' myriad decisions to remove or keep content on their platforms, following Kate Klonick's wise practice. "Regulate" is not limited to government action here, especially since, as Klonick argues, OSPs now govern (or regulate). Klonick, *supra* note 5, at 1601.

PART I

Substantive Standards

To regulate behavior for collective benefit and to diminish the social damage that it causes, it is best to define the behavior(s) in question clearly and to identify the harm that regulation is intended to prevent. OSPs have done significant work to diminish harmful content recently, responding to pressure from governments and the public. Their moderation systems are deeply flawed, however. Rules are imprecise¹¹ and inconsistently enforced.¹² Enforcement is largely limited to two reactive methods—removing content and removing accounts—which constitute a blunt instrument that has little chance of achieving durable improvement by means of behavior change, i.e., diminishing the rate at which new harmful content is posted. Removing or “taking down” content (in industry parlance) is a necessary and important tool for content moderation, but is insufficient on its own.

Finally, most of the companies govern largely in secret. They make and implement their rules with only scant input from the people whose self-expression and access to information they restrict.¹³

11 DAVID KAYE, *SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET* (2019) about the Twitter Rules: “It’s a vast and open-ended set of proscriptions.”

12 David Kaye, [Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression](#), United Nations, April 6, 2018, at 10.

13 Two significant exceptions to this are Reddit and the Wikimedia Foundation, which use what Robyn Caplan calls “the community-reliant approach”: the company sets some high-level rules as a baseline, but relies on volunteers (who vastly outnumber those companies’ employees) to both enforce its rules and establish additional norms

A. Identifying Forms of Harmful Content and Harms

There are many forms of damaging content online, and they inflict almost as many types of harm, from causing emotional distress to inspiring mass murder. To be effective, regulation of harmful online content must therefore be both clear and complex. The following list of types of harmful online content gives a sense of its variety:¹⁴

- “Hate speech”
- Celebration of terrorist acts or violence
- Content designed to recruit extremists or terrorists
- Content to organize extremists or terrorists
- Credible threats of violence
- Graphic depictions of violence
- Fake accounts/impersonation
- Incitement to violence
- Instructions for making or using weapons of mass violence
- Dangerous speech¹⁵

and guidelines for various segments of the sites. See ROBYN CAPLAN, [CONTENT OR CONTEXT MODERATION?](#) (2018). And in May 2020, the video game streaming platform Twitch established a new advisory council, half of whose members are active streamers on Twitch. See Adam Smith, [Twitch Launches Safety Advisory Council to Help Clean Up Its Platform](#), THE INDEPENDENT, May 15, 2020.

¹⁴ Scholars and researchers have developed several taxonomies of harmful online content. OSPs' publicly available rules list types of harmful content in order to prohibit them, though most companies maintain more detailed taxonomies for internal use. For some examples see, e.g., Women's Media Center, [Online Abuse 101: Internet and Jurisdiction Policy Network](#), CONTENT AND JURISDICTION PROGRAM OPERATIONAL APPROACHES 20–26 (2019); Facebook, [Community Standards](#).

¹⁵ “Dangerous speech,” my own coinage, is any form of expression (speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group. For details, including reasons why this category is useful, see [dangerousspeech.org](#).

- Bullying
- Harassment
- Abetting/promoting self-harm or suicide
- Sexual exploitation of children
- Nonconsensual or unsolicited pornography
- Defamation
- Doxing¹⁶
- Disinformation and deepfakes¹⁷
- Incitement to hatred of an identity group, which often includes falsehoods

It can be difficult to classify content into even these relatively granular categories, for several reasons. First, some of the categories (like the last two) overlap. Also, some content is not exclusively harmful: its presence online may also be constructive or beneficial. For example, human rights activists post video recordings of graphic police violence to denounce such conduct, in the hope of diminishing it,¹⁸ and law-enforcement agencies gather useful intelligence from some terrorist content.¹⁹ Also,

¹⁶ The term "doxing," derived from the word "documents" and its abbreviation "docs," means posting individuals' private information online, to expose them to harassment and attack by others.

¹⁷ Deepfakes are AI-generated videos or images that purport to show events or statements that never happened. They can be extremely difficult to identify as false. The word is a portmanteau of "deep learning" and "fake."

¹⁸ Jillian C. York, *Companies Must Be Accountable to All Users: The Story of Egyptian Activist Wael Abbas*, ELECTRONIC FRONTIER FOUNDATION (February 13, 2018); David Uberti, *How Smartphone Video Changes Coverage of Police Abuse*, COLUMBIA JOURNALISM REVIEW, April 9, 2015.

¹⁹ But see JESSICA STERN & J. M. BERGER, *ISIS: THE STATE OF TERROR* 140 (2016). They argue that in most cases the intelligence to be gathered is not valuable enough to justify allowing terrorist content to remain online.

some painful content has historic or artistic value, such as the famous 1972 photograph by Nick Ut of Phan Thi Kim Phuc, a nine-year-old Vietnamese girl who was running naked while napalm from an airstrike burned into her back and side. When a Norwegian writer, Tom Egeland, posted it in 2016 as one of “seven photographs that changed the history of warfare,” Facebook removed the image under the company’s policy against nudity. That decision elicited protests by prominent Norwegian politicians, journalists, the Norwegian prime minister, Facebook users around the world, and Kim Phuc herself, who survived the burns and now lives in Canada. Finally conceding that the historical importance of the photograph outweighed the harm of depicting a naked child in this specific case, Facebook reversed its decision.²⁰

Proposal 1

In order to prevent harm more effectively, companies should classify harmful online material not only by its content, but also by the harm it engenders. They should explain to users which forms of harm they seek to prevent, and to what degree.

As Facebook’s decisions regarding the photograph of Kim Phuc demonstrate, many key content moderation decisions are ultimately based (or *should* be based) not only on the content itself, but on the likely effects of its presence online; that is, on estimating harms and balancing them against possible benefits.

All systems of regulation, including bodies of law, are based on decisions about which harms should be suppressed and which can be tolerated.

²⁰ Sam Levin, Julia Carrie Wong, & Luke Harding, *Facebook Backs Down from “Napalm Girl” Censorship and Reinstates Photo*, THE GUARDIAN, September 9, 2016.

Companies' decisions on harm-balancing are consequential for billions of people who use their platforms. They should be both explicit and transparent, since people are more likely to follow rules whose purpose they understand.²¹ Companies should explain which harms they have chosen *not* to tolerate and why, and which content seems to produce those harms. Equally important, they should explain to users which harms they have chosen not to try to prevent and why.

Only a few companies explain their moderation policies in terms of what damage they seek to forestall, and even then, in a limited way.

Facebook, for example, gives the following policy rationale: "We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence."²² It does not mention other harms such as emotional distress, or decreased participation in civic life and discourse,²³ so users cannot know whether Facebook considers these tolerable, doesn't believe they are real, or simply chose not to mention them. Twitter says it bans what it calls "hateful conduct" (a narrower category than "hate speech," a term it does not use) because that content can curb the freedom of expression of those it denigrates; that is, it can "silence the voices of those who have been historically marginalized."²⁴ Twitter mentions no other harm. YouTube gives no public rationale for its hate-speech policy.²⁵

21 See, e.g., M. E. Tankard & E. L. Paluck, *Norm Perception as a Vehicle for Social Change*, 10.1 SOCIAL ISSUES AND POLICY REVIEW 181 (2016).

22 Facebook, *supra* note 14, at sec. 11, "Hate Speech."

23 JEREMY WALDRON, *THE HARM IN HATE SPEECH* 5 (2014): "Hate speech is both a calculated affront to the dignity of vulnerable members of society and a calculated assault on the public good of inclusiveness."

24 Twitter Inc., [Hateful Conduct Policy](#), TWITTER HELP CENTER.

25 Google, [Hate Speech Policy](#), YOUTUBE HELP.

Another reason to link policies with harms is that many forms of content inflict more than one type of harm, and often they can best be prevented with entirely different methods. For example, racist, antisemitic, or terrorist recruitment content can be deeply distressing to many people, and attractive or convincing to others. The former harm can be prevented by hiding the content—as users can do for themselves on some platforms, by means of filtering or blocking software. Removal on its own is not sufficient to prevent the latter harm, because the same recruiting material can invariably be found somewhere else online. It is worth trying and testing other methods, such as pointing users who seem to be vulnerable to recruitment toward content designed to steer them away from hatred or extremism. The eponymous Redirect Method²⁶ is one such effort.

One more reason to classify content by the harms that it may cause is that not all harms should be eliminated, even if it were possible. A significant degree of offensiveness, for example, should be tolerated to protect freedom of expression, especially political speech.

It would be interesting to discover, too, whether every rule prohibiting a type of online content can be linked to a particular harm or harms that such content seems to engender among other users of a platform. It is possible that some rules are simply normative commitments by a company's leaders and not related to any harm. If so, this too should be made explicit.

26 [Redirect Method](#). See also Lydia Dishman, *Google Algorithms and Human Psychology: How Jigsaw Rescues Teens from ISIS Recruiters*, FAST COMPANY, January 28, 2019.

As noted above, harms are nearly as varied as damaging content. Here are some examples:²⁷

- Exploitation of children
- Mental or emotional distress (caused by content not related to the viewer)
- Mental or emotional distress caused by a targeted or personal attack
- Fear of being personally assaulted, due to a credible threat
- Increased likelihood of self-harm
- Violation of privacy
- Damage to personal reputation
- Economic harm to individuals or groups (e.g., job loss)
- Silencing (decreased participation in online discourse)
- Diminished participation in civic and public life²⁸
- Increased tendency to hate or fear, discriminate against, or endorse violence against other people
- Deterioration of the tone of online discourse
- Normalization of violence and other harmful offline behavior
- Convincing people of falsehoods
- Collective harms enumerated in Article 19(3)(b) of the International Covenant on Civil and Political Rights: damage to national security, public order, public health, or morals

Finally, to make their efforts to diminish harms more effective, companies should consider classifying harms by severity or gravity. This would allow them to build triage systems and to focus on responding first, or most quickly, to the worst examples.

²⁷ For another taxonomy of harm caused by online content, *see, e.g.*, Women's Media Center, *supra* note 13; Maeve Duggan, *Online Harassment 2017*, PEW RESEARCH CENTER, July 11, 2017.

²⁸ WALDRON, *supra* note 23.

B. “Hate Speech”

The term most often used by the public, government officials, and academics to describe harmful online content is “hate speech.” In spite of its wide use there is no consensus—in law, OSP rules, or colloquial use²⁹—about what falls into that category, except egregious examples. For many of those, moreover, the term “hate speech” is not necessary, since there are other speech acts (such as incitement to violence) that are similarly defined in multiple bodies of law. As Andrew Sellars observed in a paper, “Defining Hate Speech,” in which he offers important and useful ideas toward a definition (but doesn’t quite propose one), “surprisingly little work appears to have been done to define the term ‘hate speech’ itself. Without a clear definition, how will scholars, analysts, and regulators know what speech should be targeted?”³⁰ Confusion over which speech to include has led to many cases of mistaken removal and failure to remove hateful content.³¹

Proposal 2

OSPs should clearly define which content they regulate, describe boundaries between what is prohibited and what is permitted, and explain how they take context into account.

29 See interview with Kenan Malik, *in* THE CONTENT AND CONTEXT OF HATE SPEECH: RETHINKING REGULATION AND RESPONSES 81 (Michael E. Herz & Péter Molnár eds., 2012): “If you look at hate speech laws across the world, there is no consistency about what constitutes hate speech.”

30 Andrew F. Sellars, [Defining Hate Speech](#) (2016), Berkman Klein Center Research Publication No. 2016–20, 4. For more detail on the variety of definitions, see Susan Benesch, [Defining and Diminishing Hate Speech](#), *in* STATE OF THE WORLD’S MINORITIES AND INDIGENOUS PEOPLES 2014, Minority Rights Group (2014), at 18. As pointed out by Article 19, a human rights NGO, there is no consensus definition of the term. See Article 19, [Hate Speech Explained: A Toolkit](#) (2015).

31 Davey Alba, [Defining “Hate Speech” Online Is an Imperfect Act](#), WIRE, August 22, 2017.

To contribute to a clearer and more uniform definition of online hate speech, this section summarizes existing OSP rules, national laws, and international human rights law. Each of these has special relevance: platforms' own rules form the basis of most moderation now underway; companies are obliged to comply with national laws wherever they operate; and international human rights law could serve as a universal basis for content moderation, as David Kaye, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, has proposed.³²

The human rights organization Article 19 notes that “[h]ate speech’ is an emotive concept which has no universally accepted definition in international human rights law.”³³ Perhaps it is the emotive nature of “hate speech” that has helped make the term so popular, despite its ambiguity. It is often used to signal the reader’s (or listener’s) outrage, as much as the author’s intent. As the writer and human rights advocate Salil Tripathi put it: “From speech that promotes hatred, hate speech has come to mean speech you hate. A nebulous term whose meaning varies from person to person, ‘hate speech’ is increasingly being used to vilify words and speech that we disagree with, and hence hate, expanding its meaning significantly from what it was meant to be—speech that encourages people to hate others.”³⁴ Even “hate” itself is somewhat ill-defined, as the legal scholar Robert Post has pointed out,³⁵ and it is not clear whether the “hate” in “hate speech” refers to the state of mind of the speaker/author, to the likely increase in hateful thoughts among a receptive audience,

32 Kaye, *supra* note 12.

33 Article 19, [Self-Regulation and "Hate Speech" on Social Media Platforms](#) 6 (2018).

34 Salil Tripathi, *Hate Speech*, SEMINAR 716, April 24, 2019.

35 Robert Post, *Hate Speech*, in *EXTREME SPEECH AND DEMOCRACY* 123 (Ivan Hare & James Weinstein eds., 2009).

to the speech's capacity to make people (those it attacks or purports to describe) feel hated—or, as Tripathi argues, to an expression of outrage or disagreement with the speech. The terms “hate” or “hatred,” where they are defined in law at all, are usually understood narrowly. For instance, Canada’s criminal code provision against the “willful promotion of hatred” must be “construed as encompassing only the most severe and deeply felt form of opprobrium,” the Canadian Supreme Court found in the landmark case of James Keegstra, a public school teacher who told his students that Jews are an evil people who “created the Holocaust to gain sympathy.”³⁶

Many would simply say, as US Supreme Court Justice Potter Stewart famously wrote about pornography, “I know it when I see it.”³⁷ But that would not provide consensus on what hate speech is, for many reasons. First, people identify it differently, according to their cultural backgrounds and normative commitments. Second, the meaning—and the dangerousness or capacity to bring about harm—of almost any putative hate speech depends on the context in which it is expressed or disseminated.³⁸ Third, people can be maddeningly inventive in expressing or fomenting hatred: often hateful content contains no slurs or telltale words, in part to evade detection,³⁹ but is still clearly understood by its intended audience and can be at least as vicious and powerful as

36 *R. v. Keegstra* [1990] 3 S.C.R. 697 (Can.) Part VII(D)(iii)(a) (Dickson, C.J.).

37 *Jacobellis v. Ohio*, 378 U.S. 184, at 197 (Stewart, J., concurring).

38 See DANGEROUS SPEECH PROJECT, *DANGEROUS SPEECH: A PRACTICAL GUIDE* 19 (2018).

39 Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, & Derek Ruths, *A Web of Hate: Tackling Hateful Speech in Online Social Spaces*, in *PROCEEDINGS OF THE FIRST WORKSHOP ON TEXT ANALYTICS FOR CYBERSECURITY AND ONLINE SAFETY* (2016). See also Sellars, *supra* note 29, 4:

When talking of hate speech, a shocking degree of the discussion – be it academic or in public discourse – looks solely to finding specific words or phrases that the observer

content that contains obviously hateful language. In fact, sometimes coded language or images serve as a kind of social glue, an in-joke that binds a group of people together. This is one reason why extremist and hate groups are heavy users of hand gestures with in-group meanings, or polysemic memes such as Pepe the Frog. Fourth, slurs are sometimes reclaimed by members of the group whom they ostensibly describe, who use them in non-offensive ways. Fifth, activists and targets of hate sometimes deliberately repeat speech in order to denounce it or call it out, and that content is often mistakenly censored.⁴⁰ It is therefore difficult to write—and harder to apply—rules prohibiting and accurately classifying “hate speech,” and even harder to detect it reliably with automated software tools (“classifiers”) or algorithms. This is vital to remember, since it is otherwise tempting to try to rely on software to detect and automatically remove “hate speech.”

The slipperiness of the term can also pose a serious threat to freedom of expression, since it makes it easy for governments to use it to prosecute their political opponents or minority groups. In Hungary, for example, where hateful speech against Roma is all too common and has led to violent attacks on members of that group, Roma have been prosecuted for “anti-Hungarian hate speech.”⁴¹ In Kazakhstan, a law against inciting religious hatred has been used to imprison atheists, human rights activists, and Muslims, in one case for reading a publicly available book. “Ablykhan Chalimbayev spent five years in a Kazakh prison for quoting a commentary on the Quran” under the law against religious hatred, as the

believes signal the presence of hate speech. Is that a sound strategy?

40 GILLESPIE, *supra* note 5, at 59.

41 Milkos Haraszti, *Foreword: Hate Speech and the Coming Death of the International Standard Before It Was Born (Complaints of a Watchdog)*, in *THE CONTENT AND CONTEXT OF HATE SPEECH: RETHINKING REGULATION AND RESPONSES* (2012).

Danish lawyer and human rights advocate Jacob Mchangama noted with concern.⁴² The term “hate speech” can also be used as a political weapon, as it was during Kenya’s 2013 presidential campaign, when some Kenyans felt that it was used to suppress debate, just when it was more necessary than ever.⁴³

There is a common thread in virtually all definitions of hate speech, which is that it denigrates or attacks people based on some kind of shared identity or membership in certain kinds of groups. Consequently, no matter how emphatically a person declares, “I hate you!” that is not hate speech if there is no reference to a group.

Laws and definitions of hate speech usually list specific types of groups or shared identities, such as ethnicity, religion, race, or nationality/national origin. Categories such as gender, age, sexual orientation, immigration status, disease, and/or disability are included in some definitions but not others. This has led to heated debates over which categories should “count.” Definitions vary also with regard to how severe a speech act must be to constitute hate speech: inciting violence against a member or members of a group; dehumanizing them; suggesting that they are inferior, dangerous, or too numerous (and therefore threatening); or insulting them in another way.

1. OSP Rules on Hate Speech

OSPs define hate speech in various ways and some choose not to ban it at all. Reddit CEO Steve Huffman explained why Reddit does not, saying

⁴² Jacob Mchangama, *The U.N. Hates Hate Speech More than It Loves Free Speech*, FOREIGN POLICY, February 28, 2019. See also Andrey Grishin, *How Kazakhstan's Anti-Extremism Blacklist Forces Activists, Bloggers and Opposition Politicians into the Shadows*, OPENDEMOCRACY, August 7, 2018.

⁴³ Patrick Gathara, *The Monsters under the House*, GATHARA'S WORLD (blog) (March 10, 2013).

“hate speech is difficult to define” and “it’s impossible to enforce consistently.”⁴⁴ Among those that do, YouTube calls it “promoting violence or hatred” and Facebook describes it as “violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation.” As discussed above, Twitter doesn’t ban hate speech. Instead it prohibits a narrower category that it calls “hateful conduct,” by which it means conduct that “promote[s] violence against or directly attack[s] or threaten[s] other people.” Twitter also separately prohibits the use of hateful imagery or symbols in a profile or header image. At each company, the rules have evolved over time, and changes often come in response to public controversies over specific pieces of content.

Consider Facebook’s announcement, a few days after the March 2019 massacre at two mosques in Christchurch, New Zealand, that it would ban expressions of white nationalism and white separatism. Facebook had previously considered those to be legitimate speech, distinguishing them from white supremacy, which it did identify as hate.⁴⁵ The Christchurch killer live-streamed the massacre on Facebook, and the recording was posted on many sites online; he also posted a “manifesto” in which he repeated the white supremacist claim that Muslim immigrants pose an existential threat to “Europeans” like himself. Facebook’s decision led commentators to wonder whether it would apply the same new criteria to other nationalists, not only white ones. As Salil Tripathi commented on Facebook’s announcement, “the arbitrariness of social media companies in deciding what goes on air and what doesn’t, is deeply troubling. [...]

⁴⁴ Shoshana Wodinsky, *Reddit CEO Says It’s “Impossible” to Consistently Enforce Hate Speech Rules*, THE VERGE (July 9, 2018).

⁴⁵ Tony Romm and Elizabeth Dwoskin, *Facebook Says It Will Now Block White-Nationalist, White-Separatist Posts*, WASHINGTON POST, March 27, 2019; *Standing against Hate*, FACEBOOK NEWSROOM (blog) (March 27, 2019).

Would it do the same for Hindu nationalist/Muslim fundamentalist pages?”⁴⁶

Companies also include different identity groups in their definitions of “hate speech,” effectively offering extra protection to certain groups but not others. Facebook and YouTube include caste, for example, and YouTube adds veteran status.⁴⁷ Finally, Facebook is unusual in describing three “tiers” of hate speech, all of which it ostensibly prohibits. The first is violent or dehumanizing speech, the second is speech claiming that a member or members of another group are inferior or deficient, and the third refers to calls to segregate or exclude. The current public rules of Facebook, YouTube, Twitter, and several other platforms regarding “hate speech” or hateful conduct are presented in an appendix to this paper, for reference and comparison.

It is vital to note that each set of public rules is only the tip of a much larger iceberg, since most companies have more than one set of rules: the publicly available ones such as Facebook’s “Community Standards,” and a much more detailed manual that moderators use to make decisions. The latter are kept secret,⁴⁸ which greatly limits the extent to which outsiders

46 Salil Tripathi, [@saliltripathi](#), March 29, 2019.

47 Facebook, [Community Standards](#), sec. 13. “Hate Speech”; Twitter Inc., [Hateful Conduct Policy](#), Twitter Help Center; Google, [Hate Speech Policy](#), YouTube Help.

48 Facebook published a more detailed version of its rules in 2018, but they were not nearly as extensive or granular as the ones used by moderators. Josh Constone, [Facebook Reveals 25 Pages of Takedown Rules for Hate Speech and More](#), TECHCRUNCH (April 24, 2018). In a few cases, parts of internal manuals for moderators have been leaked. See, e.g., Julia Angwin and Hannes Grassegger, [Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children](#), PROPUBLICA (June 28, 2017); Nick Hopkins, [Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence](#), THE GUARDIAN, May 21, 2017.

can understand and critique the companies' actual governance of "hate speech" and other content—the way they define such content in practice.

Since those manuals make granular distinctions between prohibited and permitted content, they should be made accessible to outsiders, including users who want to understand the details, not merely the general standards.⁴⁹ The moderators' manuals give detailed instructions for applying the standards to specific, real cases, the way regulations are used to interpret statutes. Companies have long resisted releasing their manuals, saying that this would allow bad actors to "game the system"—to find ways of remaining just barely on the permissible side of a rule, or, more generally, ways of posting vicious or harmful content while avoiding takedown. These justifications are not persuasive for two reasons. First, laws constantly draw lines between prohibited and permitted behavior, and a line is drawn in a particular way because behavior anywhere on the permitted side of the line is considered acceptable. If it is not, the line should be moved. Second, users who are determined to post harmful content and evade removal can extrapolate where the lines are, by testing the system with a variety of posts from a variety of accounts. This is commonly done, for example, by coordinated propagandists in Myanmar, according to Michael Lwin, co-founder and managing director of Koe Koe Tech, a Yangon-based IT firm.⁵⁰

Finally, OSPs should explain how they account for varied social, cultural, and political contexts when they make takedown decisions. The same hateful remark or frightening rumor can have a dramatically different capacity to influence people (and even catalyze action) in different contexts. Platforms like Facebook claim to use only one set of moderation rules for the entire world (or the large proportion of it in which they

⁴⁹ See Klonick, *supra* note 5, at 1631, distinguishing between standards and rules.

⁵⁰ Interview with the author, 2020.

operate). Surely the rules prescribe different decisions as context changes, however. This, too, should be explained for those who are governed not so much by the “Community Standards” as by their tangible application to millions of pieces of content.

2. National Laws on Hate Speech

Most bodies of national law do not mention the term “hate speech” at all, much less define it. Instead, some refer to speech acts such as incitement and discrimination—or unique to Rwanda, the vaguely and broadly defined offense of “ethnic divisionism.”⁵¹ Other laws focus on a variety of harmful consequences of speech, including insult, offence, humiliation, and degradation. Laws also identify unlawful speech by the intent of the speaker, the likely effect of the speech, and whether the speech calls for action of some kind.

US federal law famously does not criminalize hate speech. In fact, it protects the right to produce almost every form of it, under the First Amendment to the US Constitution.⁵² Only a very small subset of what would be considered hate speech by some definitions is criminalized, under the standard developed by the US Supreme Court in the 1969 case of *Brandenburg v. Ohio*. Speech can be criminal if it is “directed to inciting or producing imminent lawless action” and is also likely to successfully

51 Immigration and Refugee Board of Canada, [Rwanda: Legislation Governing Divisionism and Its Impact on Political Parties, the Media, Civil Society and Individuals](#) (2007), RWA102565.E.

52 US Const. Amend. I. “Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the government for a redress of grievances.”

incite or produce such action.⁵³ In other words, only incitement to violence that is likely to succeed quickly is prohibited. Thus US law protects “hate speech” more than any other body of law in the world—and has been highly influential in the development of OSPs’ moderation systems, since it was a formative influence on the people who designed them. As Kate Klonick observed, “American lawyers trained and acculturated in American free speech norms and First Amendment law oversaw the development of company content moderation policy. Though they might not have ‘directly imported First Amendment doctrine,’ the normative background in free speech had a direct impact on how they structured their policies.”⁵⁴

Other bodies of national law criminalize large swaths of the same speech that the US First Amendment protects. For example, §135a of the Norwegian penal code defines “hate speech” very broadly, in terms of both prohibited actions and protected identity groups. Hate speech is defined as “threatening or insulting anyone, or inciting hatred or persecution of or contempt for anyone because of his or her (a) skin color or national or ethnic origin, (b) religion or life stance, or (c) homosexuality, lifestyle or orientation.”⁵⁵ South Africa’s hate speech law is one of the most detailed and comprehensive, specifying groups and attributes that are not found in other countries’ legislation, such as pregnancy, marital status, conscience, language, skin color, and “any other group where discrimination based on that other ground (i) causes or perpetuates systemic disadvantage; (ii) undermines human dignity; or (iii) adversely

53 *Brandenburg v. Ohio*, 395 US 444 (1969).

54 Klonick, *supra* note 5.

55 *The General Civil Penal Code* (Act No. 10 of May 22, 1902, as last amended by Act No. 131, December 21, 2005), University of Oslo Law Library Translated Norwegian Legislation online database.

affects the equal enjoyment of a person's rights and freedoms in a serious manner that is comparable to discrimination [...]."⁵⁶

Bhikhu Parekh has illustrated the diversity of national laws with a set of examples:

Britain bans abusive, insulting, and threatening speech. Denmark and Canada prohibit speech that is insulting and degrading; and India and Israel ban speech that incites racial and religious hatred and is likely to stir up hostility between groups. In the Netherlands, it is a criminal offence to express publicly views insulting to groups of persons. Australia prohibits speech that offends, insults, humiliates, or intimidates individuals or groups, and some of its states have laws banning racial vilification. Germany goes further, banning speech that violates the dignity of an individual, implies that he or she is an inferior being, or maliciously degrades or defames a group.⁵⁷

Germany also prohibits denying the Holocaust in a manner that could disturb the public peace⁵⁸ and prohibits disturbing "the public peace in a

56 [Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000](#), c. 1.

57 Bhikhu Parekh, *Is There a Case for Banning Hate Speech*, in *THE CONTENT AND CONTEXT OF HATE SPEECH* 37 (Michael Herz & Péter Molnár eds., 2012). Britain's prohibition on "insulting" speech was criticized for being too broad (especially after it was used for dubious prosecutions such as one of a university student for insulting a policeman's horse) and was removed by the Crime and Courts Act 2013.

58 [German Criminal Code, Section 130\(3\)](#):

Whosoever publicly or in a meeting approves of, denies or downplays an act committed under the rule of National Socialism of the kind indicated in section 6 (1) of the Code of International Criminal Law, in a manner capable of

manner that violates the dignity of the victims by approving of, glorifying, or justifying National Socialist rule of arbitrary force.”⁵⁹

In sum, national laws on hate speech and related content vary greatly. Many of them are vague or broad enough to be difficult for OSPs to interpret, and to be subject to easy misuse by governments.⁶⁰

3. International Human Rights Law on Speech

In a 2018 report to the UN Secretary General, David Kaye, the Special Rapporteur on Freedom of Opinion and Expression, proposed that international human rights law serve as uniform guidelines for national laws on online content moderation. Private companies’ rules have created “unstable, unpredictable, and unsafe environments,” Kaye wrote. Human rights standards could be improved by the provision of “a framework for holding both States and companies accountable to users across national borders.”⁶¹ Article 19, an international freedom-of-expression organization, has made the same recommendation, arguing, like Kaye, that this would lead to clearer and more consistent rules, greater transparency about what the rules are and how they are applied, and increased opportunities for oversight.⁶²

I agree, with a caveat, that human rights law on speech is confusing and not always applicable to private companies. If properly interpreted and

disturbing the public peace shall be liable to imprisonment not exceeding five years or a fine.

⁵⁹ *Id.*, Section 130(4).

⁶⁰ Kaye, *supra* note 12, at 9. “The commitment to legal compliance can be complicated when relevant State law is vague, subject to varying interpretations or inconsistent with human rights law.”

⁶¹ *Id.*, 14.

⁶² Article 19, *supra* note 33.

explained by experts, however, it could serve as an important source of standards for content moderation by companies.

Proposal 3

International human rights law on speech can serve as a source of unified standards for moderation of "hate speech" and other harmful content by companies – after it has been analyzed and interpreted for this purpose by outside experts.

Such interpretation is particularly needed regarding hate speech, because that term is nearly absent from international law⁶³ and is not mentioned in the applicable core treaties and declarations, which refer instead to offensive, inciting, or discriminatory speech. Article 7 of the Universal Declaration of Human Rights states that all persons are entitled to protection against discrimination in violation of the Declaration—and against “any incitement to such discrimination.”⁶⁴

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) confers the right to freedom of expression and opinion. It also establishes that a state may prohibit expression only if the prohibition is: (1) provided by law, (2) necessary in a democratic society, and (3) in pursuit of one of the following aims: respect of the rights or reputations of others; or the protection of national security, public order, or public health or morals.⁶⁵ For this provision to be applied to OSPs, their rules must be understood as law. Indeed, Kaye and other scholars refer to the companies’ own

63 Hate speech makes an appearance in international criminal law as a form of persecution, which, when sufficiently widespread and systematic, can constitute a crime against humanity.

64 Universal Declaration of Human Rights, December 10, 1948, G.A. Res 217 A(III), U.S. Doc A/810 at 71 (1948).

65 International Covenant on Civil and Political Rights, December 16, 1966, G.A. Res. 2200A(XXI), 999 U.N.T.S. 171.

rules as “platform law,”⁶⁶ since they are used for governance.⁶⁷ How should the other terms be understood and used by companies? Should they make decisions about the national security of countries around the world? Societies’ public order? Morals? If so, shouldn’t they consult with stakeholders in the relevant countries? Which ones, then, and on what terms? These questions need to be answered before international human rights law can offer standards for content moderation by companies, other than in vague and general terms.⁶⁸

After Article 19 sets out the circumstances in which governments may prohibit expression, Article 20 sketches the types of expression that they must prohibit: “Any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.”⁶⁹ This provision is unclear, in no small part because the distinctions between advocacy and incitement on the one hand, and between hatred and hostility on the other, are unclear, within and among bodies of law. Jacob Mchangama has described how the odd and confusing formulation of Article 20 emerged from its contentious drafting history.⁷⁰

66 Orly Lobel, *The Law of the Platform*, 101 MINNESOTA LAW REVIEW 86 (2016), cited in Kaye, *supra* note 12.

67 Klonick, *supra* note 5; GILLESPIE, *supra* note 5.

68 Evelyn Aswad has explained how much of Article 19 of the ICCPR could be used by OSPs. See Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE LAW & TECHNOLOGY REVIEW 26–70 (2018). I am currently writing an article that offers additional ideas. See also Article 19, [Side-stepping Rights: Regulating Speech by Contract](#) (2018); United Nations, General Assembly, [Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General](#), A/74/486, October 9, 2019.

69 International Covenant on Civil and Political Rights, *supra* note 65.

70 Jacob Mchangama, *The Sordid Origin of Hate Speech Laws*, POLICY REVIEW, December 1, 2011.

Article 20 has been incorporated into bodies of national law only partially, or not at all. In light of the confusion it engenders, in 2011 and 2012 the United Nations High Commissioner for Human Rights oversaw an effort to clarify it.⁷¹ This led to the Rabat Plan of Action,⁷² which proposes a six-part threshold test for unlawful incitement. Because the test consists of factors that are often very difficult to determine online, such as the speaker's intent, it may be of limited applicability to content moderation.

Another core international human rights treaty is directly relevant to hate speech, although it omits the term. The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)⁷³ calls on its parties to "declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin."⁷⁴ This is evidently a lower threshold than the ICCPR's, and would require restricting much more speech.

Moreover, such treaties set standards that are quite general, whereas content moderation requires highly specific, granular rules. This is especially true as moderation is conducted on a large scale and OSPs need to train thousands of moderators to make consistent decisions.

71 United Nations, "[Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement to Discrimination, Hostility or Violence](#)".

72 United Nations, "[Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement to Discrimination](#)" (January 11, 2013).

73 UN General Assembly Resolution 2106A(XX), December 21, 1965.

74 UN General Assembly, [International Convention on the Elimination of All Forms of Racial Discrimination](#), UNITED NATIONS, TREATY SERIES, vol. 660, December 21, 1965.

If international human rights standards come to guide online content moderation, different platforms may derive quite different rules from them. How wide should the range of variability be?

Such questions should be resolved by experts, perhaps organized as an international council that would interpret international human rights law as it applies to content moderation by private companies. This group might be convened by the relevant UN special rapporteurs. Once international human rights law is explicated for use in private online content moderation, it can provide a useful set of universal standards.

Councils of outside advisors should not be composed only of human rights lawyers. For their recommendations to be feasible and realistic, they should include people with significant knowledge of how social media and other platforms work from the technical point of view, such as engineers, designers, and user experience (UX) researchers. In other words, the analysis of human rights law is only one area in which OSPs should seek and rely on guidance from non-governmental outsiders, including the people governed by their rules—their users.

PART II

Procedural Standards

Proposal 4

OSPs should develop councils of non-governmental outsiders to review and advise them on their content moderation rules, on both broad (national or international) and local, granular levels. In this way, their detailed rules can be properly adapted to cultural contexts, as long as this "margin of appreciation" does not lead to violations of international human rights law standards.

There is a growing consensus, now even including Mark Zuckerberg,⁷⁵ that OSPs should not write and apply rules entirely on their own. By doing so, they have operated without an external check on their rules and deprived users of agency in the basis of governance. That produces, as David Kaye puts it, a "democratic deficit."⁷⁶ "The companies, as private stewards of public space," he writes, "interfere with the idea that their users are engaging in democratic culture. Users become subjects. In that sense, platform 'life' diminishes democratic culture even as it expands the possibilities of communication."⁷⁷ Without knowledge of the rules governing online spaces, and without any sense of representation in the making of those rules, people are less likely to obey rules against hate speech and other forms of harmful behavior.

75 Zuckerberg, *supra* note 6.

76 Kaye, *supra* note 12.

77 *Id.*

Independent external review and oversight of OSPs' rules could well lead to better, more consistent regulation of harmful content online. This will require some bold experiments. First of all, many questions present themselves, such as who exactly should contribute to the rulemaking and rule-enforcing processes, how those people will be chosen, how much authority they will have, and how they will be held accountable.

Until recently, most OSPs formed only limited advisory bodies⁷⁸ such as Twitter's Trust and Safety Council, which includes online safety and anti-hatred advocates and several researchers, including from my organization. The council has no decision-making power at all. Its members simply give intermittent advice at the request of Twitter staff; sometimes the council learns of major policy changes only when Twitter announces them publicly.⁷⁹ Facebook has a Safety Advisory Board that includes some of the same members and plays a similar role, again without authority.⁸⁰

In May 2020, however, Facebook took a significant new step and created an Oversight Board to "review Facebook's most challenging content decisions—focusing on important and disputed cases."⁸¹ Notably, the board will have power to override Facebook's moderation decisions, thus shouldering responsibility for difficult cases. The board will not have the authority to review the rules themselves, but only individual decisions in which the rules were applied to remove content; nor will it have the capacity to review more than a tiny fraction of Facebook's millions of

78 As noted above, Wikimedia, Reddit, and Twitch are welcome exceptions to this. See *supra* note 13.

79 Louise Matsakis, *Twitter Trust and Safety Advisers Say They're Being Ignored*, WIRED, August 23, 2019.

80 Facebook, *What Is the Facebook Safety Advisory Board and What Does This Board Do?*

81 Facebook, *Oversight Board Charter, September 2019*.

weekly takedown decisions.⁸² The Oversight Board Charter does allow the board to issue policy recommendations—independently or at Facebook’s request—and obligates Facebook to respond publicly within 30 days.⁸³ Board members might also choose to state their opinion about the rules in their written explanations of board decisions.

Article 19 has proposed the establishment of an international social media council (or national councils) with a significantly broader ambit than that of Facebook’s Oversight Board: “The Council could elaborate ethical standards specific to the online distribution of content and cover topics such as terms and conditions, community guidelines, and the content regulation practices of social media companies.”⁸⁴ As Article 19 envisions the council, it might advise multiple OSPs.⁸⁵

In Germany, OSPs are now invited by law to consult outsiders regarding content moderation, where the purpose is to comply with German law. The Network Enforcement Act of 2017⁸⁶ (known as NetzDG, from its abbreviated German name, *Netzwerkdurchsetzungsgesetz*) is often referred to as Germany’s hate speech law, although it does not in fact prohibit hate speech. The law requires OSPs to remove content within 24 hours if it is “manifestly unlawful” under any of 22 provisions of the German penal code. It allows them to recruit independent advisors, usually lawyers, to help them “self-regulate”: to make decisions that

82 *Id.* See also Article 19, [Facebook Oversight Board: Recommendations for Human Rights-Focused Oversight](#), March 29, 2019.

83 Facebook, *supra* note 8181, at 8.

84 Article 19, [Self-Regulation and "Hate Speech" on Social Media Platforms](#) (2018), at 20.

85 *Id.* at 21.

86 [Act to Improve Enforcement of the Law in Social Networks \(Network Enforcement Act\)](#), Bundesministerium der Justiz und für Verbraucherschutz, July 12, 2017.

comply with the law in difficult cases. “Under these NetzDG partnerships, committees consisting of three lawyers will provide a legal opinion on the content they receive within seven days. Tech companies will continue to do most takedowns by themselves. The partnership committees would only receive about 5–10 ‘high-profile’ cases per month.”⁸⁷

In my view, international advisors cannot adequately contend with hate speech and other harmful forms of content, because they necessarily lack knowledge of the relevant social and political context and cannot be representative of the relevant users. OSPs should therefore recruit users to contribute both to rulemaking and to rule enforcement, at the national or even local level. This is essential for properly handling hate speech, since so much of that content can be properly understood only by those who know the detailed social, linguistic, and political context in which hate speech is made or spread. Local advisors (like their national or international counterparts) would need training in how platforms function technically, and in how to adjudicate. It would also be important, in forming local or national advisory bodies, to avoid “capture by ill-intentioned governments or groups,” as Kaye points out.

National or local bodies would be able to guide platforms in adapting enforcement of their rules to their cultural and political contexts, and in tweaking the rules to conform to local social norms (as long as those do not violate international human rights law). Most OSPs insist that they maintain a single, uniform set of rules for the world (or for all countries in which they operate)—which ostensibly means they enforce the same rule against the depiction of nudity in Sweden and in Saudi Arabia. This further distances users and their own norms from the companies’ rules, which should instead be adaptable to some extent.

87 ECHIKSON & KNOTT, *supra* note 9.

Proposal 5

OSPs should test an array of techniques for enforcing platform rules, not only removing content and accounts.

Most efforts to diminish hateful content online use only one technique: removing it or removing the accounts from which it was posted. Takedown, as it is known in the industry, is essential for some types of egregious and/or illegal content such as child sexual exploitation, but in general it is only a stopgap, and a losing game at that, since new content is posted at a staggering rate. Moreover, removing content after it is posted is reactive, not preventive. It is roughly like pursuing food safety by removing harmful food from the market, without preventing new cases of adulteration or poisoning.⁸⁸

The problem of “hate speech” online should be seen not simply as a matter of enforcing law or rules, but as a challenge to public welfare that requires behavior change, namely, building norms of tolerance and civility.

Removing content before it is posted is a tempting alternative. Some platforms, like YouTube, already automatically detect violative content and remove it immediately after it is posted; from October to December 2019, YouTube removed roughly 3.4 million videos before a single user had watched them.⁸⁹ This poses a problem for two reasons: First, hate speech cannot be automatically detected without a large margin of error.⁹⁰ Second, such removals, tantamount to prior censorship, are such strong and speech-repressive measures that they should be undertaken

⁸⁸ J. Nathan Matias, *A Toxic Web: What the Victorians Can Teach Us about Online Abuse*, THE GUARDIAN, April 18, 2016.

⁸⁹ Google Transparency Report, [Sec. YouTube Community Guidelines Enforcement](#).

⁹⁰ Saleem et al., *supra* note 39; *see also* Alba, *supra* note 31.

only with the greatest of caution, if at all, and with robust oversight from experts outside the companies that might practice it.

Already, OSPs are removing millions of posts in order to enforce their own rules, but users continue to post harmful content faster than the companies take it down. To catch up without resorting to massive automatic removal of hate speech, which could severely impinge on freedom of expression, because hate speech is so difficult to detect reliably, companies need methods to persuade people not to post harmful content in the first place.

Some internet companies and researchers have begun to test and study alternate methods. These rely on an important but overlooked insight: that not all those who produce hateful content are extremists or incorrigible “trolls.” Some are occasional offenders who behave better offline, and may be susceptible to online interventions. Alternative methods of enforcement for those who are not chronic producers of hate speech can include preventing users from posting for specific periods of time after they break a rule (some companies, such as Twitter, already do this), or requiring them to take a short online course on the rules against “hate speech.”

Widespread alarm about vicious content online should be channeled into new opportunities to define and reinforce norms of discourse and to learn, by means of rigorous research, how to influence behavior. This is not unrealistic: public concern has helped drive major behavioral change to protect people from harm, such as the wearing of seat belts in motor vehicles or the decline in smoking. Even though some people continue to transgress such norms, the majority has become compliant, keeping themselves and others safer. Norms for online discourse can be greatly improved, even without eliminating hate speech, if such norms are embraced by a critical mass of people.

Chronic offenders should be tackled differently, of course: with criminal law and prosecution where relevant, and with muscular enforcement of

platform rules. Here, too, there are options that have not been sufficiently explored, such as preventing offenders from monetizing “hate speech” and other harmful or offensive content (as YouTube does⁹¹), setting limits on both organic and paid sharing of content (WhatsApp has tried limiting the number of groups to which one user can share a piece of content⁹²), or withdrawing users’ control of online spaces. For example, some Facebook pages and YouTube channels have become highly influential, with hundreds of thousands of followers, and the users who control them as administrators can remove any comments they don’t like. Where user/administrators use this privilege to promote appalling views to a large number of followers⁹³ and to suppress dissent by anyone else, the platforms could rescind their power to do so.

On some platforms, users also have significant tools for controlling their own experience and keeping out content they don’t want to see, sometimes thanks to applications built by third-party developers, and sometimes using features that the companies provide. Facebook, for example, allows users to hide posts that contain certain keywords,⁹⁴ and Twitter’s options to “mute” and “block” accounts can be augmented by third-party tools that allow users to share their lists of blocked accounts with others.⁹⁵

91 Google, [Advertiser-Friendly Content Guidelines](#), YouTube Help.

92 WhatsApp Inc., [More Changes to Forwarding](#), WhatsApp Blog (January 21, 2019).

93 For example, pages run by extreme anti-Muslim monks in Myanmar; see Christina Fink, [Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar](#), COLUMBIA JOURNAL OF INTERNATIONAL AFFAIRS, September 17, 2018.

94 Shruthi Muraleedharan, [Keyword Snooze: A New Way to Help Control Your News Feed](#), FACEBOOK NEWSROOM (June 27, 2019).

95 Block Together: A Web App Intended to Help Cope with Harassment and Abuse on Twitter; see [Block Together](#).

The best responses for countering harmful speech online will be tailored, as much as possible, to types of content, to the audiences they reach, and to the social, cultural, and historical circumstances in which they circulate. When platforms try new methods, they should rigorously test their effects and publish the results.

Proposal 6

OSPs should communicate their rules to users more clearly and more effectively.

For almost everyone outside the companies that make and apply them, platform rules are arcane and obscure. This precludes even the possibility of basic features of democratic governance: that people take part in debating the rules, revising them, adapting them to fit their own normative or cultural contexts, defining categorical boundaries of prohibited content, and explaining the rules to others.⁹⁶ Many of these practices would be difficult to implement on massive social platforms as they are now constructed, but that's no reason to preclude them. Internet platforms will evolve and be replaced by other models. Even on existing platforms, some scholars are testing intriguing methods, such as Jenny Fan and Amy Zhang's "digital juries," to allow users to participate in governance.⁹⁷

⁹⁶ OSPs debate and revise rules internally, of course, and sometimes use the language of democratic process to describe their efforts, such as the twice-monthly "mini legislative sessions" of Facebook staff, described by Monika Bickert, vice president of consumer operations. Conference notes on file with the author; see also Alexis Madrigal, *Inside Facebook's Fast-Growing Content-Moderation Effort*, THE ATLANTIC, February 7, 2018.

⁹⁷ Jenny Fan and Amy X. Zhang, *Digital Juries: A Civics-Oriented Approach to Platform Governance*, PROCEEDINGS OF THE 2020 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (April 2020).

In the meantime, it is a dangerous precedent for most of the world to become habituated to largely invisible systems of private censorship. Moreover, making the systems more visible is not as difficult as it may seem. It would yield a variety of benefits and can be accomplished without producing collateral harm. As Tarleton Gillespie argues, “articulating the rules is the clearest opportunity for the platforms to justify their moderation efforts as legitimate.”⁹⁸

Disclosure of the rules, together with explanations of how they are applied to user requests for takedown, can also provide a sense of procedural justice that is now sorely lacking. Users of social-media platforms often complain that when they report objectionable content, the response they receive from platforms says only that their request has been denied or accepted, without reference to any particular rule.⁹⁹

There is also considerable evidence that people who are familiar with rules are more likely to follow them.¹⁰⁰ Since OSP rules are designed to prevent or at least discourage a variety of serious individual and collective harms, it would be of major social benefit if fewer internet users broke the rules and/or did so less often.

OSPs can easily make more users aware of their outward-facing content regulations. They typically present those rules in thousands of words of fine print, buried in their terms of service, which the vast majority of users never read. Many do not even know they exist.¹⁰¹

98 GILLESPIE, *supra* note 5, at 45.

99 In response to such complaints, some OSPs, including YouTube and Twitter, have begun to explain their decisions whether or not to remove content in response to individual requests for takedown.

100 See, e.g., Tankard & Paluck, *supra* note 21.

101 Anecdotal surveys by the author, in which US college students blinked in confusion when asked if they had ever read the community

In a 2017 study, every one of 543 college students in a laboratory experiment clicked the “Join” button for a new social network, unwittingly consenting in paragraph 2.3.1 of the terms of service to give the network not only their data but also their future first-born child.¹⁰² In an observational study of online behavior, fewer than 0.2% of online software buyers spent even one second looking at the terms of service before accepting them.¹⁰³ For users of OSPs, accepting these terms is, effectively, a contract of adhesion in which content-moderation rules are buried.

Writing the rules in clear, simple language and obliging users to read them is a mild and uncomplicated intervention that is very unlikely to do any harm, and can make people more likely to follow the rules. Prof. J. Nathan Matias worked with moderators on the large subreddit *r/science* to test for this effect. When the rules of the subreddit were pinned to the top of each comment thread, those who commented were significantly less likely to break the rules.¹⁰⁴

Other efforts to improve online norms of behavior by making rules visible give some early basis for cautious optimism. There have been reports of successful online behavior modification: by Facebook, to teach users to resolve grievances successfully with one another;¹⁰⁵ by the online gaming

guidelines or content rules of any platform they used. Many students said they were not aware that such rules existed, or were available to read.

102 David Berreby, *Click to Agree with What? No One Reads Terms of Service, Studies Confirm*, THE GUARDIAN, March 3, 2017.

103 Yannis Bakos, Florencia Marotta-Wurgler, & David R. Trossen, *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts*, 43.1 J. LEGAL STUD. 1 (2014).

104 J. Nathan Matias, “[Governing Human and Machine Behavior in an Experimenting Society](#)” (2017) (unpublished PhD dissertation, Massachusetts Institute of Technology).

105 Jason Marsh, *Can Science Make Facebook More Compassionate?*, GREATER GOOD MAGAZINE, July 25, 2012.

company Riot Games, to decrease “toxic” comments by players of League of Legends, a game played by millions around the world;¹⁰⁶ and even as far back as the 1990s, by the Massachusetts Institute of Technology’s then-director of academic computing, to reduce online harassment of students.¹⁰⁷ These experiences should provide a trove of information, but thus far the findings have not been published in sufficient detail to permit replication or statistical analysis. It is essential to build up an accessible and rigorous body of knowledge about ways to diminish harmful online behavior, e.g., by communicating the rules clearly.

When they began the discussion site Parlio in 2014 with the goal of fostering civil public conversation among people who strongly disagree with one another, Wael Ghonim and his co-founders required new users to read and accept a simple set of rules, presented one at a time in a relatively large font and few words, so it was almost impossible to ignore them. So many users remained civil that the platform staff found itself with very few moderation dilemmas to discuss at their weekly meetings. (They did note that their users may have been a disproportionately civil sample of the population even before they joined Parlio.¹⁰⁸) Parlio also posted a brief statement emphasizing the new site’s focus on—and enforced demand for—civility.¹⁰⁹

106 Brendan Maher, *Can a Video Game Company Tame Toxic Behavior?* NATURE, March 30, 2016.

107 Gregory A. Jackson, *Promoting Network Civility at MIT: Crime & Punishment, or the Golden Rule?*, 75.3 EDUCATIONAL RECORD 29 (1994).

108 Interview on file with author, 2015.

109 The text of the statement:

Be curious, open-minded, and civil. We want you to share opinions and experiences that strengthen the community’s collective intelligence. We believe diversity of thought is a virtue, and we’re here to learn new perspectives; not to win arguments. We are trying to define a new type of network. One

Of course, not all users will be swayed by such interventions. Many will continue to ignore rules, some will be unable to understand them, and highly motivated trolls and other producers of harmful content may even be inspired to work harder to flout them. Some of those producers are not only highly motivated but vigorously supported and/or employed by governments in many countries, such as Russia, China, and Brazil.¹¹⁰ However, there is evidence that, at least on some platforms, a majority of the hateful content is produced not by chronic trolls or bad actors but by users who do so only occasionally.

In internal research at the company Riot Games, Jeffrey Lin found that only about 1% of League of Legends players were consistently producing what he called toxic content, and that they were responsible for less than 5% of such content on the platform.¹¹¹ The rest was produced by intermittent violators who were usually civil. If a critical mass of those users becomes familiar with the rules, there may be a net favorable effect as the rules become better accepted as robust norms of behavior.

It is an old and familiar process, after all: many of the major improvements in human life in recent decades are the result of behavior change driven by shifts in social norms, such as not smoking, wearing seatbelts, boiling unsanitary water before giving it to infants, and so on. Though some people fail to comply and some vigorously continue to violate norms, most enjoy both the individual and collective benefits. Further, in this case making the rules transparent may plant seeds for other forms of effective engagement by users.

void of Internet-trolling, where we can create a community of trust and respect that expands our horizons. Parlio values dissent, but above all else, civility.

110 SAMUEL C. WOOLLEY & PHILIP N. HOWARD, [COMPUTATIONAL PROPAGANDA: POLITICAL PARTIES, POLITICIANS, AND POLITICAL MANIPULATION ON SOCIAL MEDIA](#) (2018).

111 Maher, *supra* note 106.

Proposal 7

OSPs should allow external oversight of enforcement mechanisms.

OSPs have been rapidly expanding, speeding up, and automating their content-moderation systems, and will continue to do so under increasing pressure from users and especially from governments. Yet no one outside the companies has anything more than a vague and anecdotal sense of which content is being taken down and which content remains online. The photograph of the Vietnamese girl running while napalm burned her body, for example, galvanized extensive public discussion about Facebook's rules and especially about how they are enforced, but it is just one piece of content out of the approximately one million such pieces that Facebook removes every day, not including content that it classifies as spam.¹¹² Though it is important to publicize rules, it is also critical to understand how they are being put into practice or enforced. In order to protect freedom of expression, it is therefore essential to construct a mechanism for oversight.

Such a mechanism could have serious implications for user privacy, since it would be difficult to review actual moderation decisions without seeing examples of real content, posted from real accounts. Among other concerns, tech companies worry that releasing such data might expose them to prosecution for failing to remove content that governments deem illegal.

User data could be protected in one of two ways. First, data could be released only under a rigorous system of the type commonly used in the social sciences for sensitive data that includes private information.

112 Facebook, [Community Standards Enforcement Report](#), Facebook Transparency Report (May 2020).

Such data can be accessed only under strictly controlled conditions, and only by researchers who have been vetted in advance. In this case, outside reviewers would be forbidden to release any actual content, or information about individual users or accounts. Second, data could be released exclusively to independent boards that would be set up for that purpose and vetted in advance.

Finally, some testing can be done even without data provided by OSPs, though researchers might then violate an OSP's terms of service, which often prohibit the "scraping" of data. Some countries also have laws against this, including the American Computer Fraud and Abuse Act, which criminalizes "unauthorized access" to a website.¹¹³ In any case, review of moderation systems should become standard procedure, just as food safety inspectors visit restaurant kitchens on an intermittent but regular basis.

113 See, e.g., Brian Z. Mund, *Comment, Protecting Deceptive Academic Research under the Computer Fraud and Abuse Act*, 37 YALE L. & POL'Y REV. 385 (2018).

OSP Hate Speech Policies

The following are excerpts from various internet companies' policies on hate speech or hateful conduct, or relevant portions of their terms of service or other similar documents. In some cases, sections not related to hateful speech have been omitted for clarity and brevity.

Facebook¹

We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion, and in some cases, may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.

Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others. In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. People sometimes express contempt in the context of a romantic break-up. Other times, they use gender-exclusive language to control membership in a health or positive support group, such as a breastfeeding group for women only. In all of these cases,

1 Facebook, [Community Standards](#), sec. 13, "Hate Speech."

we allow the content but expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.

We allow humor and social commentary related to these topics. In addition, we believe that people are more responsible when they share this kind of commentary using their authentic identity.

Do not post:

Tier 1

Content targeting a person or group of people (including all subsets except those described as having carried out violent crimes or sexual offences) on the basis of their aforementioned protected characteristic(s) or immigration status with:

- Violent speech or support in written or visual form
- Dehumanizing speech such as reference or comparison to:
 - Insects
 - Animals that are culturally perceived as intellectually or physically inferior
 - Filth, bacteria, disease and feces
 - Sexual predator
 - Subhumanity
 - Violent and sexual criminals
 - Other criminals (including but not limited to "thieves," "bank robbers," or saying "all [protected characteristic or quasi-protected characteristic] are 'criminals'")
- Mocking the concept, events or victims of hate crimes, even if no real person is depicted in an image
- Designated dehumanizing comparisons in both written and visual form

Tier 2

Content targeting a person or group of people on the basis of their protected characteristic(s) with:

- Generalizations that state inferiority (in written or visual form) in the following ways:

- Physical deficiencies are defined as those about:

- Hygiene, including but not limited to: filthy, dirty, smelly
- Physical appearance, including but not limited to: ugly, hideous

- Mental deficiencies are defined as those about:

- Intellectual capacity, including but not limited to: dumb, stupid, idiots
- Education, including but not limited to: illiterate, uneducated
- Mental health, including but not limited to: mentally ill, retarded, crazy, insane

- Moral deficiencies are defined as those about:

- Culturally perceived negative character trait, including but not limited to: coward, liar, arrogant, ignorant
- Derogatory terms related to sexual activity, including but not limited to: whore, slut, perverts

- Other statements of inferiority, which we define as:

- Expressions about being less than adequate, including but not limited to: worthless, useless

- Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females."

- Expressions about deviating from the norm, including but not limited to: freaks, abnormal
- Expressions of contempt or their visual equivalent, which we define as:
 - Self-admission to intolerance on the basis of a protected characteristic, including but not limited to: homophobic, islamophobic, racist
 - Expressions that a protected characteristic shouldn't exist
 - Expressions of hate, including but not limited to: despise, hate
- Expressions of dismissal, including but not limited to: don't respect, don't like, don't care for
- Expressions of disgust or their visual equivalent, which we define as:
 - Expressions that suggest the target causes sickness, including but not limited to: vomit, throw up
 - Expressions of repulsion or distaste, including but not limited to: vile, disgusting, yuck
- Cursing, such as:
 - Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole
 - Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker
 - Terms or phrases calling for engagement in sexual activity, or contact with genitalia or anus, or with feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit

Tier 3

Content targeting a person or group of people on the basis of their protected characteristic(s) with any of the following:

- Calls for segregation
- Explicit exclusion, which includes, but is not limited to, "expel" or "not allowed."
- Political exclusion defined as denial of right to political participation.
- Economic exclusion defined as denial of access to economic entitlements and limiting participation in the labor market,
- Social exclusion defined as including, but not limited to, denial of opportunity to gain access to spaces (incl. online) and social services.

We do allow criticism of immigration policies and arguments for restricting those policies.

Content that describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics.

Twitter²

Hateful conduct

You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

2 [Hateful Conduct Policy](#), Twitter Help Center.

Hateful imagery and display names

You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

Rationale

Twitter’s mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right—we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives.

We recognize that if people experience abuse on Twitter, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. This includes; women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature and have a higher impact on those targeted.

We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals with abuse based on protected category.

If you see something on Twitter that you believe violates our hateful conduct policy, please report it to us.

When this applies

We will review and take action against reports of accounts targeting an individual or group of people with any of the following behavior, whether within Tweets or Direct Messages.

Violent threats

We prohibit content that makes violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., “I will kill you.”

Note: we have a zero tolerance policy against violent threats. Those deemed to be sharing violent threats will face immediate and permanent suspension of their account.

Wishing, hoping or calling for serious harm on a person or group of people

We prohibit content that wishes, hopes, promotes, or expresses a desire for death, serious and lasting bodily harm, or serious disease against an entire protected category and/or individuals who may be members of that category. This includes, but is not limited to:

- Hoping that someone dies as a result of a serious disease, e.g., “I hope you get cancer and die.”
- Wishing for someone to fall victim to a serious accident, e.g., “I wish that you would get run over by a car next time you run your mouth.”
- Saying that a group of individuals deserve serious physical injury, e.g., “If this group of protesters don’t shut up, they deserve to be shot.”

References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims

We prohibit targeting individuals with content that references forms of violence or violent events where a protected category was the primary target or victims, where the intent is to harass. This includes, but is not limited to sending someone:

- media that depicts victims of the Holocaust;
- media that depicts lynchings.

Inciting fear about a protected category

We prohibit targeting individuals with content intended to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., “all [religious group] are terrorists.”

Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone

We prohibit targeting individuals with repeated slurs, tropes or other content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes about a protected category. This includes targeted misgendering or deadnaming of transgender individuals.

We also prohibit the dehumanization of a group of people based on their religion, age, disability, or serious disease.

Hateful imagery

We consider hateful imagery to be logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, gender identity or ethnicity/national origin. Some examples of hateful imagery include, but are not limited to:

- symbols historically associated with hate groups, e.g., the Nazi swastika;
- images depicting others as less than human, or altered to include hateful symbols, e.g., altering images of individuals to include animalistic features; or
- images altered to include hateful symbols or references to a mass murder that targeted a protected category, e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust.

Media depicting hateful imagery is not permitted within live video, account bio, profile or header images. All other instances must be marked as sensitive media. Additionally, sending an individual unsolicited hateful imagery is a violation of our abusive behavior policy.

Do I need to be the target of this content for it to be a violation of the Twitter Rules?

Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.

When we review this type of content, it may not be clear whether the intention is to abuse an individual on the basis of their protected status, or if it is part of a consensual conversation. To help our teams understand the context, we sometimes need to hear directly from the person being targeted to ensure that we have the information needed prior to taking any enforcement action.

Note: individuals do not need to be a member of a specific protected category for us to take action. We will never ask people to prove or disprove membership in any protected category and we will not investigate this information.

Consequences

Under this policy, we take action against behavior that targets individuals or an entire protected category with hateful conduct, as described above. Targeting can happen in a number of ways, for example, mentions, including a photo of an individual, referring to someone by their full name, etc.

When determining the penalty for violating this policy, we consider a number of factors including, but not limited to the severity of the violation

and an individual's previous record of rule violations. For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent account suspension. If an account is engaging primarily in abusive behavior, or is deemed to have shared a violent threat, we will permanently suspend the account upon initial review.

YouTube³

Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes:

- Age
- Caste
- Disability
- Ethnicity
- Gender Identity and Expression
- Nationality
- Race
- Immigration Status
- Religion
- Sex/Gender
- Sexual Orientation
- Victims of a major violent event and their kin
- Veteran Status

3 [Hate Speech Policy](#), YouTube Help.

If you see content that violates this policy, please report it. If you have found multiple videos, comments, or a user's entire channel that you wish to report, please visit our reporting tool, where you will be able to submit a more detailed complaint.

What this means for you

If you're posting content

Don't post content on YouTube if the purpose of that content is to do one or more of the following.

- Encourage violence against individuals or groups based on any of the attributes noted above. We don't allow threats on YouTube, and we treat implied calls for violence as real threats. You can learn more about our policies on threats and harassment.
- Incite hatred against individuals or groups based on any of the attributes noted above.

Other types of content that violates this policy

- Dehumanizing individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease, or any other non-human entity.
- Praise or glorify violence against individuals or groups based on the attributes noted above.
- Use of racial, religious or other slurs and stereotypes that incite or promote hatred based on any of the attributes noted above. This can take the form of speech, text, or imagery promoting these stereotypes or treating them as factual.
- Claim that individuals or groups are physically or mentally inferior, deficient, or diseased based on any of the attributes noted above. This

includes statements that one group is less than another, calling them less intelligent, less capable, or damaged.

- Allege the superiority of a group over those with any of the attributes noted above to justify violence, discrimination, segregation, or exclusion.
- Conspiracy theories ascribing evil, corrupt, or malicious intent to individuals or groups based on any of the attributes noted above.
- Call for the subjugation or domination over individuals or groups based on any of the attributes noted above.
- Deny that a well-documented, violent event took place.
- Attacks on a person's emotional, romantic and/or sexual attraction to another person.
- Content containing hateful supremacist propaganda including the recruitment of new members or requests for financial support for their ideology.
- Music videos promoting hateful supremacism in the lyrics, metadata, or imagery.

Educational content

We may allow content that includes hate speech if the primary purpose is educational, documentary, scientific, or artistic in nature. This is not a free pass to promote hate speech. Examples include:

- A documentary about a hate group: Educational content that isn't supporting the group or promoting ideas would be allowed. A documentary promoting violence or hatred wouldn't be allowed.
- A documentary about the scientific study of humans: A documentary about how theories have changed over time, even if it includes theories about the inferiority or superiority of specific groups, would be allowed

because it's educational. We won't allow a documentary claiming there is scientific evidence today that an individual or group is inferior or subhuman.

- Historical footage of an event, like WWII, which doesn't promote violence or hatred.

This policy applies to videos, video descriptions, comments, live streams, and any other YouTube product or feature. For educational content that includes hate speech, this context must appear in the images or audio of the video itself. Providing it in the title or description is insufficient.

Examples

Here are examples of hate speech not allowed on YouTube.

- "I'm glad this [violent event] happened. They got what they deserved [referring to persons with the attributes noted above]."
- "[Person with attributes noted above] are dogs" or "[person with attributes noted above] are like animals."

More examples

- "Get out there and punch a [person with attributes noted above]"
- "Everyone in [groups with attributes noted above] are all criminals and thugs."
- "[Person with attributes noted above] is scum of the earth."
- "[People with attributes noted above] are a disease."
- "[People with attributes noted above] are less intelligent than us because their brains are smaller."
- "[Group with any of the attributes noted above] threaten our existence, so we should drive them out at every chance we get."

- “[Group with any of the attributes noted above] has an agenda to run the world and get rid of us.”
- “[Attribute noted above] is just a form of mental illness that needs to be cured.”
- “[Person with any of the attributes noted above] shouldn’t be educated in schools because they shouldn’t be educated at all.”
- “All of the so-called victims of this violent event are actors. No one was hurt, and this is just a false flag.”
- “All of the ‘so-called victims’ of this are actors. No one was hurt.”
- Shouting “[people with attributes noted above] are pests!” at someone regardless of whether the person does or does not have the alleged attributes
- Video game content which has been developed or modified (“modded”) to promote violence or hatred against a group with any of the attributes noted above.

Please remember these are just some examples, and don’t post content if you think it might violate this policy.

What happens when content violates this policy

If your content violates this policy, we’ll remove the content and send you an email to let you know. If this is your first time violating our Community Guidelines, you’ll get a warning with no penalty to your channel. If it’s not, we’ll issue a strike against your channel. If you get 3 strikes, your channel will be terminated.

If we think your content comes close to hate speech, we may limit YouTube features available for that content.

Instagram⁴

Instagram is a reflection of our diverse community of cultures, ages, and beliefs. We've spent a lot of time thinking about the different points of view that create a safe and open environment for everyone.

We created the Community Guidelines so you can help us foster and protect this amazing community. By using Instagram, you agree to these guidelines and our Terms of Use. We're committed to these guidelines and we hope you are too. Overstepping these boundaries may result in deleted content, disabled accounts, or other restrictions.

Follow the law.

Instagram is not a place to support or praise terrorism, organized crime, or hate groups.

Respect other members of the Instagram community.

We want to foster a positive, diverse community. We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages. We do generally allow stronger conversation around people who are featured in the news or have a large public audience due to their profession or chosen activities.

It's never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your intent clearly.

⁴ [Community Guidelines](#), Instagram Help Center.

Serious threats of harm to public and personal safety aren't allowed. This includes specific threats of physical harm as well as threats of theft, vandalism, and other financial harm. We carefully review reports of threats and consider many things when determining whether a threat is credible.

Be thoughtful when posting newsworthy events.

We understand that many people use Instagram to share important and newsworthy events. Some of these issues can involve graphic images. Because so many different people and age groups use Instagram, we may remove videos of intense, graphic violence to make sure Instagram stays appropriate for everyone.

We understand that people often share this kind of content to condemn, raise awareness or educate. If you do share content for these reasons, we encourage you to caption your photo with a warning about graphic violence. Sharing graphic images for sadistic pleasure or to glorify violence is never allowed.

Tumblr⁵

What Tumblr is for:

Tumblr celebrates creativity. We want you to express yourself freely and use Tumblr to reflect who you are, and what you love, think, and stand for.

What Tumblr is not for:

- **Terrorism.** We don't tolerate content that promotes, encourages, or incites acts of terrorism. That includes content which supports or celebrates terrorist organizations, their leaders, or associated violent activities.

5 Tumblr, [Community Guidelines](#), January 23, 2020.

- **Hate Speech.** Don't encourage violence or hatred. Don't post content for the purpose of promoting or inciting the hatred of, or dehumanizing, individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, veteran status, sexual orientation, disability or disease. If you encounter content that violates our hate speech policies, please report it.

Keep in mind that a post might be mean, tasteless, or offensive without necessarily encouraging violence or hatred. In cases like that, you can always block the person who made the post—or, if you're up for it, you can express your concerns to them directly, or use Tumblr to speak up, challenge ideas, raise awareness or generate discussion and debate.

- **Violent Content and Threats, Gore and Mutilation.** Don't post content which includes violent threats toward individuals or groups—this includes threats of theft, property damage, or financial harm. Don't post violent content or gore just to be shocking. Don't showcase the mutilation or torture of human beings, animals (including bestiality), or their remains. Don't post content that encourages or incites violence, or glorifies acts of violence or the perpetrators.

- **Harassment.** Don't engage in targeted abuse, bullying, or harassment. Don't engage in the unwanted sexualization or sexual harassment of others. If someone is sending you unwanted messages, or reblogging your posts in an abusive way, we encourage you to be proactive. Report them, and block the hell out of them. And if someone blocks you, don't attempt to circumvent the block feature or otherwise try to communicate with them.

If we conclude that you are violating these guidelines, you may receive a notice via email. If you don't explain or correct your behavior, we may take action against your account. We do our best to ensure fair outcomes, but in all cases we reserve the right to suspend accounts, or remove content, without notice, for any reason, but particularly to protect our services,

infrastructure, users, and community. We reserve the right to enforce, or not enforce, these guidelines in our sole discretion, and these guidelines don't create a duty or contractual obligation for us to act in any particular manner.

Microsoft

Microsoft Services Agreement⁶

3. Code of Conduct.

a. By agreeing to these Terms, you're agreeing that, when using the Services, you will follow these rules:

...

vii. Don't engage in activity that is harmful to you, the Services, or others (e.g., transmitting viruses, stalking, posting terrorist or violent extremist content, communicating hate speech, or advocating violence against others).

b. **Enforcement.** If you violate these Terms, we may stop providing Services to you or we may close your Microsoft account. We may also block delivery of a communication (like email, file sharing or instant message) to or from the Services in an effort to enforce these Terms or we may remove or refuse to publish Your Content for any reason. When investigating alleged violations of these Terms, Microsoft reserves the right to review Your Content in order to resolve the issue. However, we cannot monitor the entire Services and make no attempt to do so.

Community Standards for Xbox⁷

We built Xbox Live for people like you—for players from all walks of life, everywhere in the world, who all want the same thing: a place to play and

⁶ Microsoft, [Microsoft Services Agreement](#), July 1, 2019.

⁷ Microsoft, [Community Standards for Xbox](#).

have fun. We need your help keeping the Xbox online community safe and fun for everyone.

While the Code of Conduct section of the Microsoft Services Agreement applies to all Microsoft products, Xbox Live offers so many ways to interact with others that it benefits from an additional level of explanation.

To this end, we've created the following community standards for Xbox. Consider these standards a roadmap for contributing to this incredible, globe-spanning community. Remember: Xbox Live is your community. We all bring something unique, and that uniqueness is worth protecting.

Whether you're brand new to gaming or have been playing for decades, we need you to be stewards of this place, to protect each other even as you compete. Because when everyone plays, we all win.

Our Shared Values

The spirit of Xbox lives in our values, which are key to sustaining a vibrant and welcoming community. Living these values every time we play shows the world the unifying power of gaming.

- Gaming can be enjoyed by all
- Creativity powers community
- Competition is best when it's fair
- Helping others makes all of us stronger
- Hate has no place here

Conduct

Some parts of the internet don't have rules—and the Xbox online community isn't one of them. Yes, Xbox Live is, in a meaningful sense, your gaming network. But it belongs to millions of others, too. You

deserve a place to be yourself with confidence, free from bullying, hatred, and harassment—and so does every other player. So it's important to treat others as they would like to be treated.

Remember:

- Win or lose, be a good sport
- Did someone have a great game? Let them know!
- You are the community
- A little bit of trash talk is okay, but keep it clean
- No one likes trolling, so don't do it

Content

The gamertags, gamerpics, screenshots, game clips, and other posts you make on Xbox can be a great way to show off what's meaningful to you. We encourage all players to be themselves and show off what they like, what makes them laugh, or what makes them amazing. But this sharing can't come at the expense of other players' positive experiences.

Remember:

- Use your skills and creativity to add informative, helpful, funny, or interesting content that contributes positively to our vibrant and diverse community
- Content you post on Xbox needs to suit a wide audience
- Context is important, and mature content that makes sense in a game might not be appropriate elsewhere on Xbox
- Not everyone has the same likes or dislikes as you, so think twice about saying something hurtful about someone else's content, playing style, or choices

Standards

If you've seen the Microsoft Services Agreement, the following rules probably look familiar. They may sound a bit like legalese, but bear with

us—upholding these standards is critical to maintaining a community where everyone can have fun! People differ about what seems fun, and conflicts sometimes occur. But while plenty of conflicts can be worked out between players, there are nevertheless some things we just can't tolerate.

In each section you'll find examples showing how the Microsoft Services Agreement's Code of Conduct relates to Xbox Live.

ii. Do your part to keep everyone safe

To keep Xbox Live a place where everyone can have fun, we can't allow behavior or content designed to exploit, harm, or threaten anyone – children, adults, or otherwise. When threatening, abusive, or insulting language is used against another member of our community, or the community at large, it undermines every player's ability to enjoy themselves.

For example, don't:

- Threaten someone with physical assault after an intense game
- Message other players with homophobic slurs
- Make a club grounded in ethnic hatred
- Create a Looking for Group that negatively calls out another player
- Post insults in another player's activity feed
- Respond to someone's smack talk with sexual slurs

iv. Keep your content clean

People enjoy all shapes and styles of content on Xbox. Everyone's tastes are different, and that's great! However, that doesn't mean that absolutely anything goes. To keep Xbox Live welcoming and inclusive for everyone, some content must be avoided.

Support a welcoming and inclusive community

Harassment and hate take many forms, but none have a home on Xbox. To make Xbox Live a place where everyone can hang out, and to prevent people from feeling uncomfortable or unwelcome, we all need to be stewards. This means more than just not harassing other players—it means embracing them. It means saving those unsavory jokes for people you know will enjoy them. It means taking particular care for others while you play, keeping in mind how they might interpret your content.

For example, don't:

- Make fun of other people's identities or personal traits
- Send harassing or abusive messages
- Use a club to shame other players or groups
- Start a broadcast in order to troll someone
- Flood voice chat with music during a multiplayer match
- Post game clips that will offend many others

Know the difference between trash talk and harassment

We get it—gaming can be competitive and interactions with other players can get heated. A little trash talk is an expected part of competitive multiplayer action, and that's not a bad thing. But *hate has no place here*, and what's not okay is when that trash talk turns into harassment.

Trash talk includes any lighthearted banter or bragging that focuses on the game at hand and encourages healthy competition. *Harassment* includes any negative behavior that's personalized, disruptive, or likely to make someone feel unwelcome or unsafe. To qualify as harassment, the behavior doesn't have to be drawn-out or persistent. Even a single abusive message could harm someone's experience. Know when to draw the line, when to back off. Know and respect the other player.

For example:

Acceptable trash talk includes

Get destroyed. Can't believe you thought you were on my level.

That was some serious potato aim. Get wrecked.

Only reason you went positive was you spent all game camping. Try again, kid.

Cheap win. Come at me when you can actually drive without running cars off the road.

That sucked. Get good and then come back when your k/d's over 1.

Going too far looks like

Get <sexual threat>. Can't believe you thought you were on my level.

Hey <profanity>, that was some serious potato aim. Get wrecked, trash.

Only reason you went positive was you spent all game camping. KYS, kid.

Cheap win. Totally expected from a <racial slur>.

You suck. Get out of my country— maybe they'll let you back in when your k/d's over 1.

Consequences

Our priority is the safety and enjoyment of everyone on Xbox Live. Content and behavior that puts players at risk or makes them feel unwelcome has no place in the Xbox online community. So, sometimes we need to step in. We're not out to punish, but rather to protect everyone's experience.

Every suspension or other corrective action aims only to show what was wrong and what can be learned from a situation. When suspensions end, we welcome players back so they can contribute to Xbox Live in positive ways. We know people make mistakes, and we believe lapses in judgment can be significant opportunities for growth.

Inappropriate conduct

If you violate Xbox community standards, you may find restrictions placed on your profile and/or device. When we suspend an Xbox profile, we restrict access to features that are most closely associated with the problematic behavior. Most commonly, this means a temporary suspension that removes one or more features for a period of time. Temporary suspensions can include:

- Restrictions on the use of online multiplayer gaming
- Removal of the ability to send text and voice messages on Xbox
- Blocking real-time voice and text communications on Xbox
- Preventing the broadcast of live game play
- Restrictions on the use of parties and clubs

Inappropriate content

Since Xbox Live content must be appropriate for all audiences, sometimes we remove content to protect our customers. Depending on the type of content violation, this can result in our restricting certain features for the profile that created or shared the content. Temporary suspensions can include:

- Blocks on the ability to upload game clips and screenshots to Xbox Live
- Restrictions on uploading or sharing Kinect content
- Removal of inappropriate content from Xbox Live
- Automatic assignment of a new gamertag
- Limits on the ability to share Xbox content on other social networks
- Removal of the ability to edit your Xbox profile or clubs

Repeat or severe offenses

We may permanently suspend a profile or device if we can no longer trust it due to a severe violation, or if our attempts to correct repeated negative

behaviors are unsuccessful. Under permanent suspension, the owner of the suspended profile forfeits all licenses for games and other content, Gold membership time, and Microsoft account balances.

Microsoft Hate Speech Reporting Form⁸

At Microsoft, we recognize that we have an important role to play in fostering safety and civility on our hosted consumer services.

Please use this web form to report content posted or shared on Microsoft-hosted consumer services that may constitute hate speech - for example, content that advocates violence or promotes hatred based on:

- Age
- Disability
- Gender
- National or ethnic origin
- Race
- Religion
- Sexual orientation
- Gender identity

Please note that not all content that you may find offensive is considered hate speech and, in reviewing your report, Microsoft may choose to take no action.

⁸ Microsoft, "[Report Hate Speech Content Posted to a Microsoft Hosted Consumer Service](#)."

WhatsApp⁹

Legal and Acceptable Use.

You must access and use our Services only for legal, authorized, and acceptable purposes. You will not use (or assist others in using) our Services in ways that:

- (b) are illegal, obscene, defamatory, threatening, intimidating, harassing, hateful, racially, or ethnically offensive, or instigate or encourage conduct that would be illegal, or otherwise inappropriate, including promoting violent crimes;
- (c) involve publishing falsehoods, misrepresentations, or misleading statements

Pinterest¹⁰

Our team works hard to keep divisive, disturbing or unsafe content off Pinterest. We delete some types of content, and other stuff we just hide from public areas.

We remove hate speech and discrimination, or groups and people that advocate either. Hate speech includes serious attacks on people based on their race, ethnicity, national origin, religion, gender identity, sexual orientation, disability or medical condition. Also, please don't target people based on their age, weight, immigration or ex-military status.

We remove content used to threaten or organize violence or support violent organizations. We don't allow anything that presents a real risk of

⁹ WhatsApp.com, [WhatsApp Terms of Service](#), January 28, 2020.

¹⁰ Pinterest, [Community Guidelines](#).

harm to people or property. We also don't want anyone making threats, organizing violence or encouraging others to be violent.

Any person or group that's dedicated to causing harm to others isn't welcome on Pinterest. That includes terrorist organizations and gangs. We collaborate with industry, government and security experts to help us identify these groups.

We remove harmful advice, content that targets individuals or protected groups and content created as part of disinformation campaigns.

Don't put harmful misinformation on Pinterest.

- We don't allow advice when it has immediate and detrimental effects on a pinner's health or on public safety. This includes promotion of false cures for terminal or chronic illnesses and anti-vaccination advice.
- We don't allow misinformation about protected groups that promotes fear, hate and prejudice. This and other policies, including our hate speech guidelines, are designed to keep Pinterest a positive and welcoming environment for people of all backgrounds.
- We don't allow misinformation that attacks individuals and turns them, their families or their properties into targets of harassment or violence.
- We don't allow content that originates from disinformation campaigns targeted at Pinterest or other platforms.
- We don't allow false or misleading content that impedes the integrity of an election or an individual's or group's civic participation, including registering to vote, voting, and being counted in a census.

Airbnb

Airbnb Community Standards: Fairness¹¹

The global Airbnb community is as diverse, unique, and vibrant as the world around us. Fairness is what holds us together, what makes it possible for us to trust one another, integrate seamlessly within communities, and feel as if we can truly belong.

Discriminatory behavior or hate speech

You should treat everyone with respect in every interaction. So, you should follow all applicable laws and not treat others differently because of their race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, disability, or serious diseases. Similarly, insulting others on these bases is not allowed.

Airbnb's Nondiscrimination Policy: Our Commitment to Inclusion and Respect

Airbnb is, at its core, an open community dedicated to bringing the world closer together by fostering meaningful, shared experiences among people from all parts of the world. Our community includes millions of people from virtually every country on the globe. It is an incredibly diverse community, drawing together individuals of different cultures, values, and norms.

The Airbnb community is committed to building a world where people from every background feel welcome and respected, no matter how far they have traveled from home. This commitment rests on two foundational principles that apply both to Airbnb's hosts and guests: inclusion and respect. Our shared commitment to these principles enables every

11 Airbnb, [Community Standards](#).

member of our community to feel welcome on the Airbnb platform no matter who they are, where they come from, how they worship, or whom they love. Airbnb recognizes that some jurisdictions permit, or require, distinctions among individuals based on factors such as national origin, gender, marital status or sexual orientation, and it does not require hosts to violate local laws or take actions that may subject them to legal liability. Airbnb will provide additional guidance and adjust this nondiscrimination policy to reflect such permissions and requirements in the jurisdictions where they exist.

While we do not believe that one company can mandate harmony among all people, we do believe that the Airbnb community can promote empathy and understanding across all cultures. We are all committed to doing everything we can to help eliminate all forms of unlawful bias, discrimination, and intolerance from our platform. We want to promote a culture within the Airbnb community—hosts, guests and people just considering whether to use our platform—that goes above and beyond mere compliance. To that end, all of us, Airbnb employees, hosts and guests alike, agree to read and act in accordance with the following policy to strengthen our community and realize our mission of ensuring that everyone can belong, and feels welcome, anywhere.

- **Inclusion** – We welcome guests of all backgrounds with authentic hospitality and open minds. Joining Airbnb, as a host or guest, means becoming part of a community of inclusion. Bias, prejudice, racism, and hatred have no place on our platform or in our community. While hosts are required to follow all applicable laws that prohibit discrimination based on such factors as race, religion, national origin, and others listed below, we commit to do more than comply with the minimum requirements established by law.

- **Respect** – We are respectful of each other in our interactions and encounters. Airbnb appreciates that local laws and cultural norms vary

around the world and expects hosts and guests to abide by local laws, and to engage with each other respectfully, even when views may not reflect their beliefs or upbringings. Airbnb's members bring to our community an incredible diversity of background experiences, beliefs, and customs. By connecting people from different backgrounds, Airbnb fosters greater understanding and appreciation for the common characteristics shared by all human beings and undermines prejudice rooted in misconception, misinformation, or misunderstanding.

Specific Guidance for Hosts in the United States and European Union

Guided by these principles, our U.S. and EU host community will follow these rules when considering potential guests and hosting guests:

Race, Color, Ethnicity, National Origin, Religion, Sexual Orientation, Gender Identity, or Marital Status

Airbnb hosts **may not**

- Decline a guest based on race, color, ethnicity, national origin, religion, sexual orientation, gender identity, or marital status.
- Impose any different terms or conditions based on race, color, ethnicity, national origin, religion, sexual orientation, gender identity, or marital status.
- Post any listing or make any statement that discourages or indicates a preference for or against any guest on account of race, color, ethnicity, national origin, religion, sexual orientation, gender identity, or marital status.

Gender Identity

Airbnb does not assign a gender identity to our users. We consider the gender of an individual to be what they identify and/or designate on their user profile.

Airbnb hosts may not

- Decline to rent to a guest based on gender unless the host shares living spaces (for example, bathroom, kitchen, or common areas) with the guest.
- Impose any different terms or conditions based on gender unless the host shares living spaces with the guest.
- Post any listing or make any statement that discourages or indicates a preference for or against any guest on account of gender, unless the host shares living spaces with the guest.

Airbnb hosts **may**

- Make a unit available to guests of the host's gender and not the other, where the host shares living spaces with the guest.

Age and Familial Status

Airbnb hosts **may not**:

- Impose any different terms or conditions or decline a reservation based on the guest's age or familial status, where prohibited by law.

Airbnb hosts **may**:

- Provide factually accurate information about their listing's features (or lack of them) that could make the listing unsafe or unsuitable for guests of a certain age or families with children or infants.
- Note in their listing applicable community restrictions (e.g. senior housing) that prohibit guests under a particular age or families with children or infants.

Disability

Airbnb hosts **may not**:

- Decline a guest based on any actual or perceived disability.
- Impose any different terms or conditions based on the fact that the guest has a disability.
- Substitute their own judgment about whether a unit meets the needs of a guest with a disability for that of the prospective guest.
- Inquire about the existence or severity of a guest's disability, or the means used to accommodate any disability. If, however, a potential guest raises his or her disability, a host may, and should, discuss with the potential guest whether the listing meets the potential guest's needs.
- Post any listing or make any statement that discourages or indicates a preference for or against any guest on account of the fact that the guest has a disability.
- Refuse to communicate with guests through accessible means that are available, including relay operators (for people with hearing impairments) and e-mail (for people with vision impairments using screen readers).
- Refuse to provide reasonable accommodations, including flexibility when guests with disabilities request modest changes in your house rules, such as bringing an assistance animal that is necessary because of the disability, or using an available parking space near the unit. When a guest requests such an accommodation, the host and the guest should engage in a dialogue to explore mutually agreeable ways to ensure the unit meets the guest's needs.

Airbnb hosts **may**:

- Provide factually accurate information about the unit's accessibility features (or lack of them), allowing for guests with disabilities to assess for themselves whether the unit is appropriate to their individual needs.

When guests are turned down. Hosts should keep in mind that no one likes to be turned down. While a host may have, and articulate, lawful and legitimate reasons for turning down a potential guest, it may cause that member of our community to feel unwelcome or excluded. Hosts should make every effort to be welcoming to guests of all backgrounds. Hosts who demonstrate a pattern of rejecting guests from a protected class (even while articulating legitimate reasons) undermine the strength of our community by making potential guests feel unwelcome, and Airbnb may suspend hosts who have demonstrated such a pattern from the Airbnb platform.

Specific Guidance for Hosts Outside the United States and European Union

Outside of the United States and the European Union, some countries or communities may allow or even require people to make accommodation distinctions based on, for example, marital status, national origin, gender or sexual orientation, in violation of our general nondiscrimination philosophy. In these cases, we do not require hosts to violate local laws, nor to accept guests that could expose the hosts to a real and demonstrable risk of arrest, or physical harm to their persons or property. Hosts who live in such areas should set out any such restriction on their ability to host particular guests in their listing, so that prospective guests are aware of the issue and Airbnb can confirm the necessity for such an action. In communicating any such restrictions, we expect hosts to use clear, factual, non-derogatory terms. Slurs and insults have no place on our platform or in our community.

What happens when a host does not comply with our policies in this area?

If a particular listing contains language contrary to this nondiscrimination policy, the host will be asked to remove the language and affirm his or her understanding and intent to comply with this policy and its underlying

principles. Airbnb may also, in its discretion, take steps up to and including suspending the host from the Airbnb platform.

If the host improperly rejects guests on the basis of protected class, or uses language demonstrating that his or her actions were motivated by factors prohibited by this policy, Airbnb will take steps to enforce this policy, up to and including suspending the host from the platform.

As the Airbnb community grows, we will continue to ensure that Airbnb's policies and practices align with our most important goal: To ensure that guests and hosts feel welcome and respected in all of their interactions using the Airbnb platform. The public, our community, and we ourselves, expect no less than this.

The Internet and social media have revolutionized the production, consumption, and flow of information around the world, allowing billions of people to interact with one another and to exchange views and ideas. Unfortunately, the same digital platforms also facilitate the circulation of harmful content, including online hate speech, and its dissemination on an unprecedented scale at remarkable speed. Addressing the harmful consequences of online hate speech poses a unique regulatory challenge, because such speech is typically hosted on private platforms that operate in digital space and outside the control of most national governments.

The present publication contains the results of a joint research project undertaken by the Israel Democracy Institute (IDI) and Yad Vashem, with the goal of supporting efforts by social media companies and other internet intermediaries to formulate policies and policy guidelines that can reduce online hate speech. The first part consists of the **IDI/Yad Vashem Recommendations for Reducing Online Hate Speech**: sixteen recommendations, developed by a team of senior international experts in the field, which are meant to serve as the basis of policy guidelines for social media companies and internet intermediaries. The second part consists of studies by members of the research team—Tehilla Shwartz-Altshuler (IDI) and Mr. Rotem Medzini (IDI), Prof. Karen Eltis (University of Ottawa) and Dr. Iliia Siatitsa (Geneva Academy of Human Rights and International Humanitarian Law), and Prof. Susan Benesch (Berkman Klein Center, Harvard University)—that analyze online platforms' current policies and the legal frameworks in which they operate, and propose directions for future reforms. The recommendations and research papers can inform the current debates about the regulation of online speech and influence the positions of the stakeholders who participate in such debates—governments, international organizations, academia, civil society, the technology sector, the media, and the public at large.

www.en.idi.org.il



0 4500001227 1
450-1227 דאגאקוד

ISBN:
978-965-519-294-0

May 2020
