Dealing with Hate Speech on Social Media

Rotem Medzini | Tehilla Shwartz Altshuler

THE ISRAEL DEMOCRACY INSTITUTE

יד ושם
YAD VASHEM
THE WORLD HOLOCAUST REMEMBRANCE CENTER

Policy
Paper
E12

THE ISRAEL DEMOCRACY INSTITUTE

יד ושם
YAD VASHEM
THE WORLD HOLOCAUST REMEMBRANCE CENTER

Policy Paper E12

DEALING WITH HATE SPEECH ON SOCIAL MEDIA

Rotem Medzini | Tehilla Shwartz Altshuler

June 2019

Text Editor: Lenn Schramm
Series and Cover Design: Studio Tamar Bar Dayan
Typesetting: Nadav Shtechman Polischuk
Printed by Maor Wallach Print

ISBN: 978-965-519-257-5

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 policy paper 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 to 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 impart 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

ABSTRACT

Though the need to prevent hate speech was not born with social-media platforms, the rise in its volume 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 do their policies also protect other groups that people are part of – whether involuntary 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 for public statements than for 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 work 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 until a final decision is taken, 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, 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.

Introduction

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 phenomenon that accompany these innovations—from disinformation and fake news, through the infringement of privacy, mass surveillance, harmful psychological side-effects, and influencing elections, and on to the accumulation of wealth and political power that results from the 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 such as Facebook, Instagram, YouTube, and Twitter. 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

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

whether Americans are more likely to be exposed to 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 to do 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 need to prevent hate speech directed against individuals and groups was not born with social-media platforms. However, the leveling of hierarchies and the 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* (03.12.2017) [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 there is no disagreement 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 online service providers (hereinafter 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 recommendations aim at permitting 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 for 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 test[12] 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 16 December 1966, entry into force 23 March 1976, in accordance with Article 49.

[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 5 October 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 direct," and the "nature of the arguments deployed in the speech";

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]

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

(5) the extent, reach, and size of the audience; (6) the likelihood that the speech will effectively incite harm (ibid.).

[13] The Convention had been 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, or ethnicity, and religion. *See* UNESCO – Countering Online Hate Speech, *supra* note 14.

[17] See other papers in this project.

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 their 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, their 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, 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

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

[19] Daphne Keller, *Litigating Platform Liability in Europe: New Human Rights Case Law in the Real World, The Center for Internet and Society* (13.04.2016).

[20] Delfi AS v. Estonia, application no. 64569/09.

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 personality 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).

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

[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).

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 comments is and 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 the hate speech and direct threats.

1.2.3. The Council of Europe's Cybercrime Convention and its Additional Protocol[23]

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]

[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 23 November 2001, entered into force on 01 July 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 signature on 28 January 2003, entered into force in 1 march 2006 (ETS no. 189).

[24] *Id.*

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 (28 November 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 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]

[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 crime, their protection, and the specific needs related to their recognition, respectful treatment, support, and access to justice.

[26] Council Framework Decision 2008/913/JHA of 28 November 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, *available at* http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3Al33178 (last visited: March 27, 2019).

[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

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]

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]

committed or who are liable for offenses motivated by racism or xenophobia, including antisemitism.

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

[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

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 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]

over online hate speech based on (7) the location of the server and (8) whether the content is accessible 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).

[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 the Press Freedom in Kantor Center – Legislation Survey, *supra* note 31, at 40.

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

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.

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 peace[40] or is an advocacy for, or promotion of genocide.[41]

[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.

[39] These exclusions include statements that are proven by the defendant to be true, statements that are offered "in good faith," 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.

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]

[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, 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, 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 are in any case base their decision 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 in order to prevent the use of their platform 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 crime, 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 about copyright infringement, the reaction to cybercrime and terrorist content, and the definition 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 a "strict liability" regime, which requires 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, Elonnai 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 extend the scope of this paper. The following paper in our project, deal 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 Roomates.com was not entitled to the CDA immunity, because its drop-down menus were structured to facilitate user's 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 to 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 internent intermediaries have no incentive to install self-monitoring mechanisms. However, when an internet intermediary is notified of illegal content and thus receive "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 of 8 June 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 to respond to requests by copyright owners who identified 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 addresses the issue of the liability for 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. Agencia de 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 (provides clarifications for OSP' liability for trademark infringement committed by their users on 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, 12 July 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 in 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 Act, 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, take-down 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, along with 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 using 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 to Moderate 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 result 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 did 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 its 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 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, it 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 relatively a 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 Schmalensee, Matchmakers: the New Economics of Multisided Platforms (2016).

[87] Code of Conduct on Countering Illegal Hate Speech Online (31.5.2016).

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 terrorists 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 the community members. The next paragraphs address these different form 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

[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.

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

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 (GIFTC), 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 company-level policies 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 conducts 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." Community Guidelines, YouTube. *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 consider "innocent" sharing of hate speech 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] For setting policies in the content 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 addition 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 fight bias elimination.

[96] Kathleen Chaykowski, *Facebook is Hiring 3,000 Moderators in Push to Curb Violate Videos,* Forbes (May 3, 2017).

[97] For instance, the 2012 abuse standards included: (1) slurs or racial comments of any kind; (2) attacking 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) showing support for organizations and people primarily known for violence; (5) depicting symbols primarily known for hate and violence, unless comments are clearly against them; (6) "versus photos" comparing two people (or

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]

an 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 with 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 to classify 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 onto 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 inductive or contraindicative, and whether the text includes hate speech or not. The bag-of-words technique offers some additional benefits on top of 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 and 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 and implicit structures in the discourse; e.g., what role

[102] For instance, we can think of the next 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 an hate speech take-down policy.

is taken by participants and therefore whether each one he 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 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 interfaces 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 a hate-crime 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 a 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 lower 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, a limit 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 (Dec. 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] For the implementation on 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* (22.09.2017).

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 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.

[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 policies[116] 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 to Define 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 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-holders. According to Article 13,

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 for 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 criteria, 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

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)) (29 June 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 on the implementation of each criterion and whether it should be implemented it 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).

Scales for content moderation (lenient to stricter options)

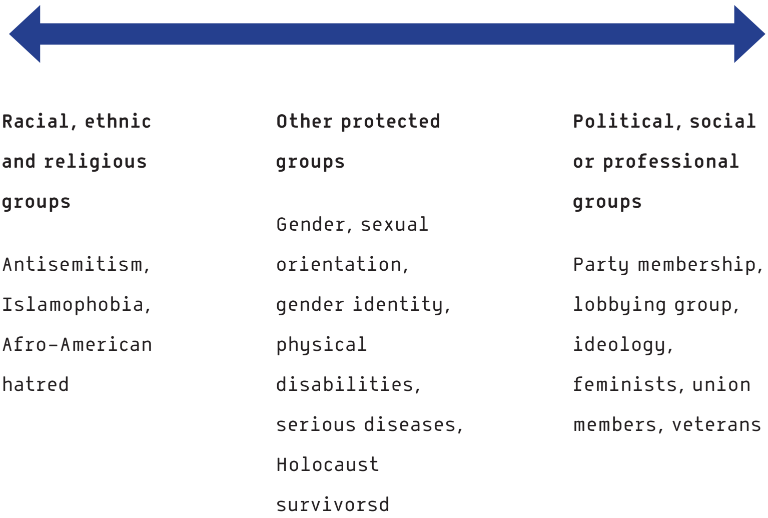
Minimizing interventions
Maximizing interventions

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	Other protected groups	Political, social or professional groups
Antisemitism, Islamophobia, Afro-American hatred	Gender, sexual orientation, gender identity, physical disabilities, serious diseases, Holocaust survivorsd	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 to discriminate against or speak 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 protected 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, an 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	Mixed approach	Context-based approach
Banned expressions, symbols, or images list	Bag of words, conjunctions, linguistic structures	Satire, historically significant event, 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, the content should 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 need to look for. The annotated texts can then be fed into predictive models that try to learn and generalize.

[124] One such list is the Wikipedia list of ethnic or religious slurs. While these list were created by the Wikipedia community, other lists could be created through a 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.

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 significant or important to the public interest. The decision

[129] Under algorithm-based instruments (starting in 5.2.5.4) 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).

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-word approaches regarding linguistic and discourse structures. These approaches include

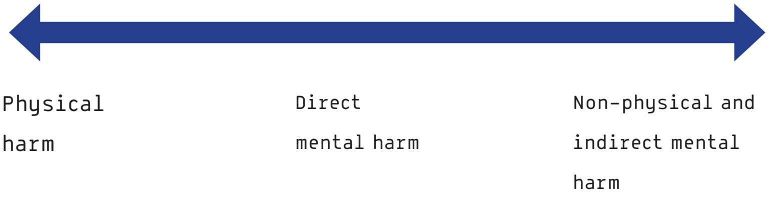
[132] Facebook's community standards mention this balance under both safety and voice. *See* Facebook Community Standards, *supra* note 93.

[133] See *The Importance of Context*, YouTube Help.

sentiment analysis to identify negative sentiments or 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]

3. 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.

Speech that causes harm



[134] For instance, following the discovery in 2018 that Facebook did not remove 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 team 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, Nov. 5, 2018. *See also* Steve Stecklow, *Why Facebook is Losing the War on Hate Speech in Myanmar,* Aug. 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.

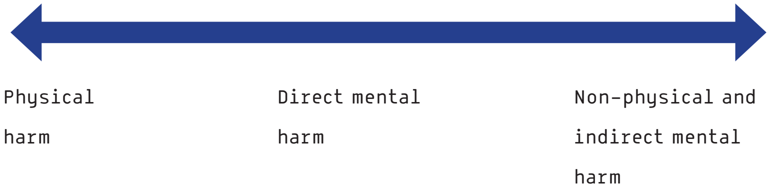
c. *Non-physical and indirect mental harm* refers to hate speech that affects and influences the target's relationships with others, financial situation, performance at work, and social and personal life. It can include a refusal to hire or rent an apartment, which we do not see as falling into the category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

3a. **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 groups, 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 (Nov. 28, 2008); Twitter Rules, Twitter.

[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.

Speech that causes harms



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 advices states to look 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).

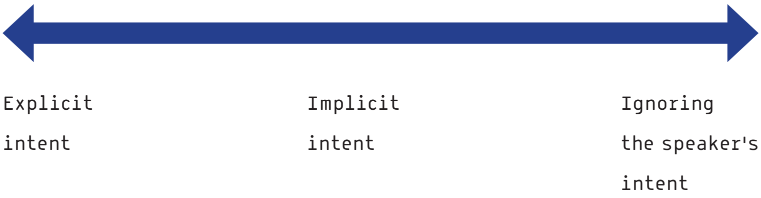
[138] The Rabat Plan of Action ¶ 22.

[139] Sellars, *supra* note 27, at 28.

[140] Facebook defines harassment as sending messages that repeat 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. 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.

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]

Intent to harm



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 from previous statements. Some NPL technologies such as sentiment analysis and linguistic structures try to tackle the problem of implicit intent. For instance, sentiment analysis can help determine out if a text expresses

[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*.

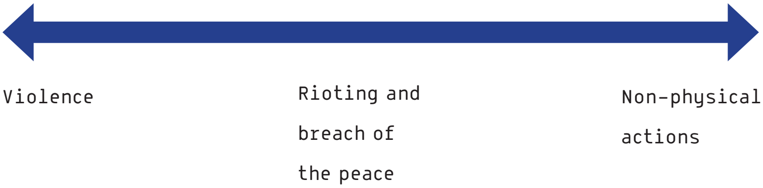
[143] *See* Sellars, *supra* note 27, at 27.

positive 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.

4. **The speech incites to socially undesirable action:** This criterion addresses a requirement that the speech may incite other consequence. In the American context, the incitement must be imminent or almost inevitable.[144]

Socially undesirable action



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 (Nov. 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 to review 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.

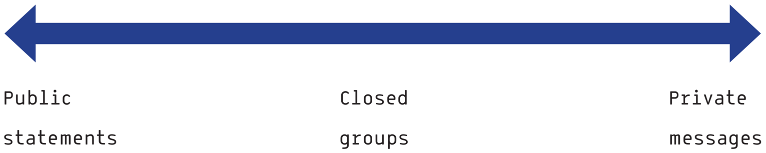
[148] Allan, *Hard Questions*, *supra* note 140.

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 others, 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 in most platforms). Current laws (as in Canada and Australia[149]) 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.

The public spread of the speech



[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 require 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 a public group.

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 so it 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

[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.

murder.[152] Some content moderators have committed 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

[152] *Facebook Failing to Protect Moderators from Mental Trauma, Lawsuit Claims,* The Guardian, September 25, 2018.

[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*.

obtain access.[156] OSPs have recently begun publishing transparency 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]

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

[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

4.3.3. Creating a user interface for submitting complaints:

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 content to help it identify the content, add voiceover or text narration to explain it, and what users should not do about the 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

Illegal Hate Speech Online: One Year After (June 2017). For more information see YouTube's Trusted Flaggers program.

[160] YouTube, Guidelines for Adding Content.

[161] According to the European Union, in 51.4% of the cases, OSPs assessed notifications in less that 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.

at the bottom of pages, hidden behind several web-clicks, or require filling in a form and copying over the address of the original 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

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

[164] Wojcicki, *supra* note 109.

or on a button next to it to access it, thus affirming their informed consent to exposure to the offensive material.

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. Based on 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

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

[166] *See id.*

law-enforcement agencies, there is no external oversight or possibility of requesting a court order.

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. Unlike content flagged by law-enforcement agencies, which 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 Law mentions the possibility of contacting the user who posted the content,[167] Twitter, because it accepts

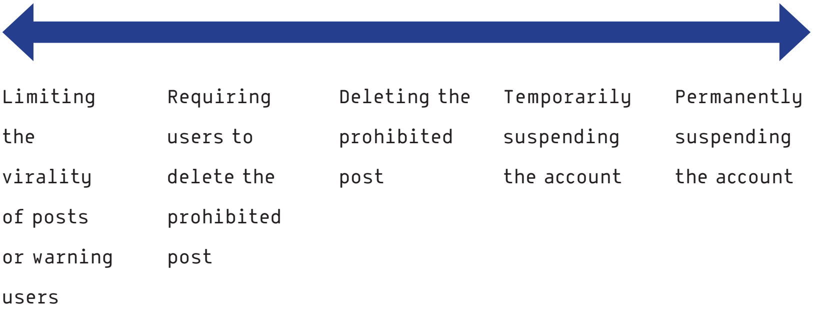
[167] *See* §3(2).3 of the German Network Enforcement Act.

reports from anyone, states that 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:

Appropriate enforcement actions



Limiting the virality of posts or warning users

Requiring users to delete the prohibited post

Deleting the prohibited post

Temporarily suspending the account

Permanently suspending the account

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

[168] Twitter Hateful Content policy.

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

[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] 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. OSPs should provide an appeal mechanism 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.

[178] *Id.*

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 external oversight or appeal mechanism for content moderation, we consider this mechanism to be highly dependent on the OSP's economic capacity and platform's 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 time 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 OSPs' policies, but the content moderation team decided to remove it from the platform.
>
> 4.5.4.2. Flagged content was found to violate the OSPs' policies, but the content moderation team decided not to remove it from the platform.

4.5.5. To assist 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 policies to implement

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-resolutions 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 states 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 and 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, and reflect 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 counties. 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 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, enables 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,* The 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 NPL 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 website.

[191] Frank Pasquale, The Black Box Society: The Secret Algorithm that Control Money and Information (2015).

[192] Anupam Chander, The Racist Algorithm?, 115 Mich. L. Rev. 1023 (2017).

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

[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), at http://fra.europa.eu/sites/default/files/fra_uploads/fra-2016-antisemitism-update-2005-2015_en.pdf. *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).

of the European member states,[199] the data collected by the European institutions cannot be compared due to gross under-reporting of the extent, nature, and characteristics of antisemitic incidents in Europe. As such, 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 race, ethnicity, 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.

[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.).

[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 race, ethnicity, 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 were directed against Jews and 22.2% against Muslims. *See id.*

One such policy was introduced after the adoption of the Code of Conduct on Countering Illegal Hate Speech Online[202] by the European 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

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

[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, "[i]n 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

there is still room for improvement in the platforms' transparency and 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

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 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).

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 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]

[211] *Id.* at 14.

[212] *Id.* at 15.

[213] *Id.* at 14–17

[214] *Id.* at 39.

[215] *Id.* at 15.

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 and web 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 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

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 to track 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

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.

[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*.

in the case of African Americans and Hispanics.[223] Pew also found that an 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 conversation) 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

[223] 68% of African American and 58% of Hispanic social-media users say that at least some 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).

[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.

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 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 a 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]

[228] *Id.*

[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.

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 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.

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 Organization.

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 separate 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.

○ 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."

She did 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,* Sept. 22, 2017).

[238] Similar to 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 Facebook 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

[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, Oct. 21, 2016, *available at* https://www.wsj.com/articles/facebook-employees-pushed-to-remove-trump-posts-as-hate-speech-1477075392 (last visited: March 27, 2019).

[242] Facebook, Community Standards: Hate Speech, https://www.facebook.com/communitystandards#hate-speech (last visited: March 27, 2019).

seeing new, violating content appearing on a Page or Facebook Profile.

■ "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 how to use report links:

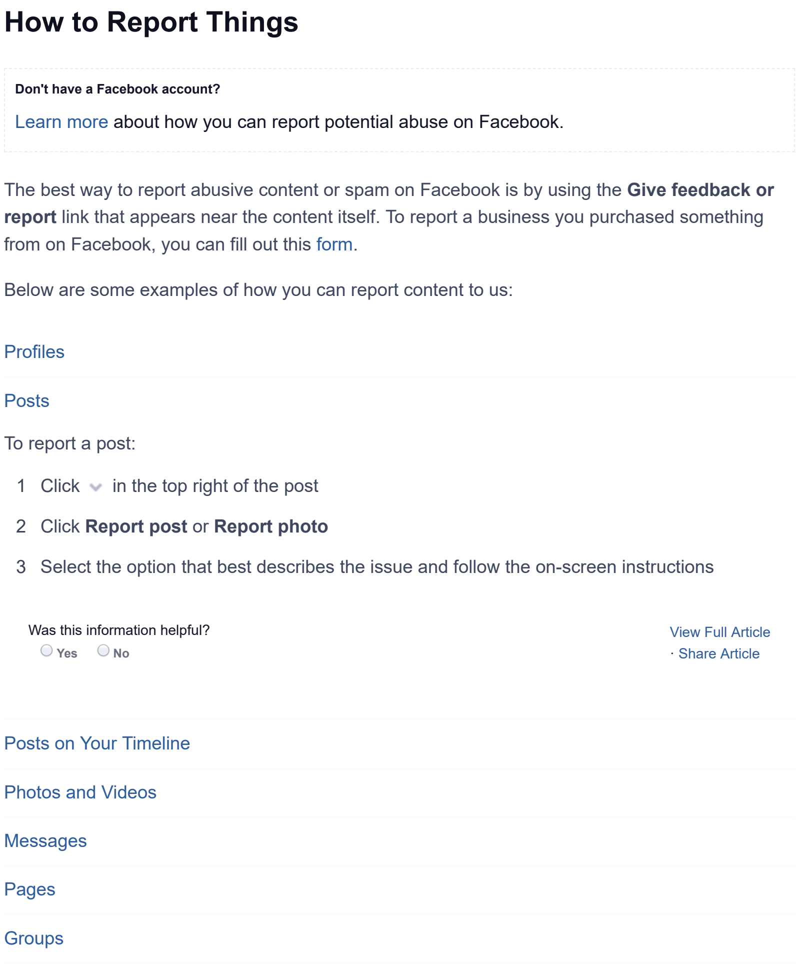
■ First, users can use "report links" to send a message to the person who posted the content and request that the content be removed.

■ If users feel uncomfortable about reaching out to the speaker directly, Facebook suggest 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.

[243] *Id.*

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 governments 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 https://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 did 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 apply 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 YouTube users to take these rules seriously. They are requested not to look for loopholes or to try to lawyer their way around the guidelines, but only to understand and respect them.

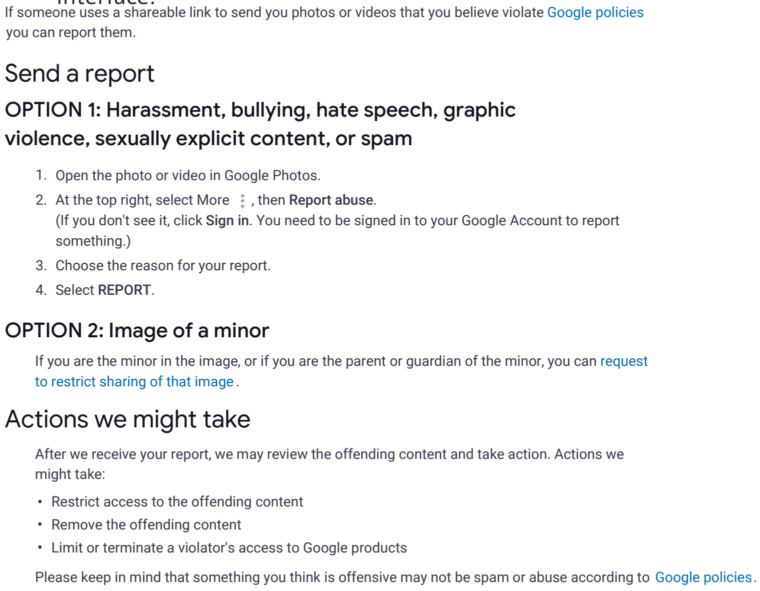
○ For hate crimes, specifically, YouTube repeats the Google definition of hate crimes. In addition, its policies add that "there is a fine line between what is and what is not considered to be hate speech. For instance, it is acceptable to criticize a nation state, but if the primary

[247] *Id.*

[248] YouTube, Policies and Safety.

purpose of the content is to incite 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.

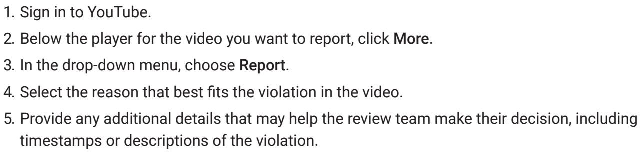
[249] Google, Prohibited and Restricted Content.

○ 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]."

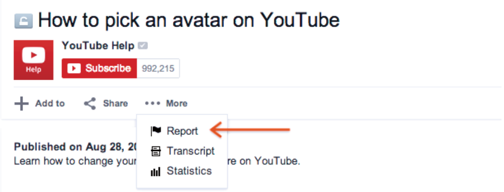
○ Flagging of videos is anonymous, so other users cannot tell who flagged a video.

○ YouTube allows users to flag videos, thumbnails, comments, live chat messages, channels, and playlists.[250]

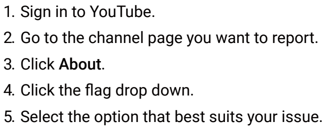
■ How to flag a video:



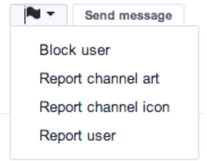
1. Sign in to YouTube.
2. Below the player for the video you want to report, click **More**.
3. In the drop-down menu, choose **Report**.
4. Select the reason that best fits the violation in the video.
5. Provide any additional details that may help the review team make their decision, including timestamps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



■ How to flag a channel:



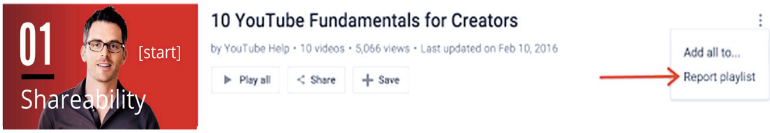
1. Sign in to YouTube.
2. Go to the channel page you want to report.
3. Click **About**.
4. Click the flag drop down.
5. Select the option that best suits your issue.



[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**

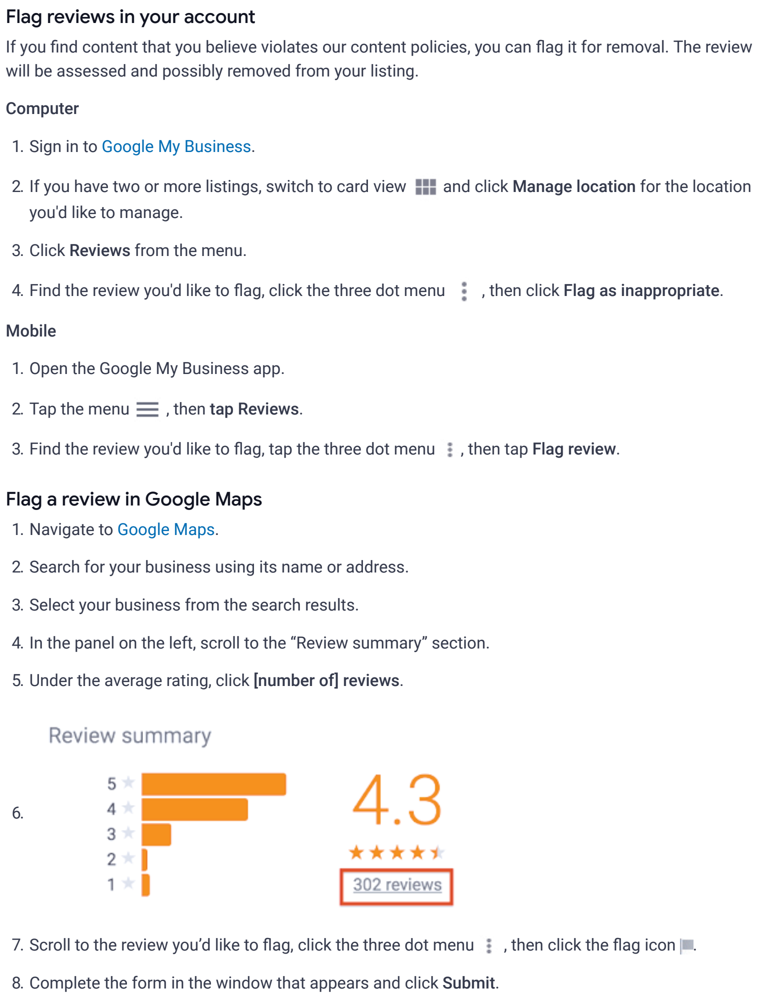

10 YouTube Fundamentals for Creators

by YouTube Help • 10 videos • 5,066 views • Last updated on Feb 10, 2016

- 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

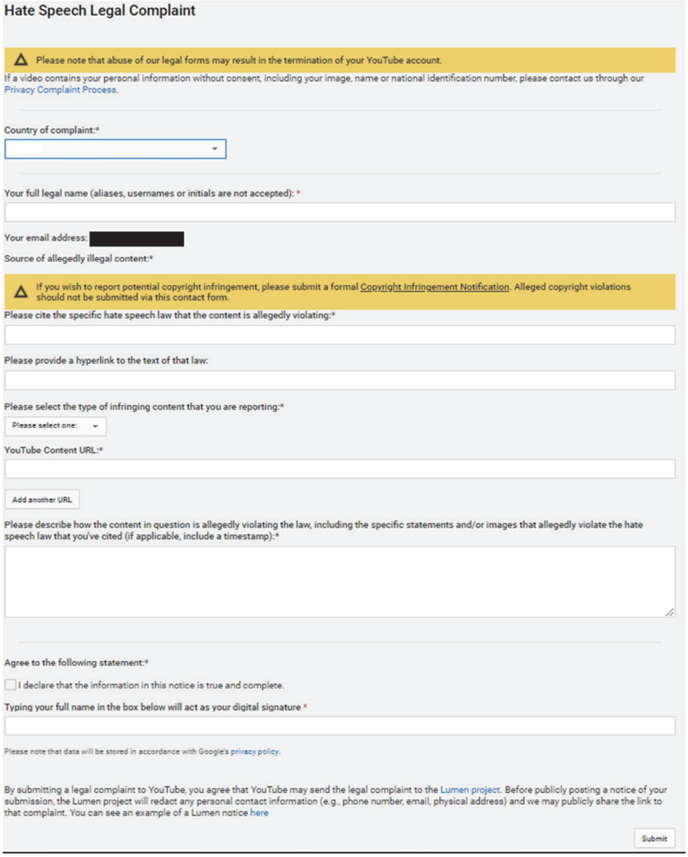
1. Navigate to Google Maps.
2. Search for your business using its name or address.
3. Select your business from the search results.
4. In the panel on the left, scroll to the “Review summary” section.
5. Under the average rating, click **[number of] reviews**.
6.

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: [redacted]

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:*

YouTube Content 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 *

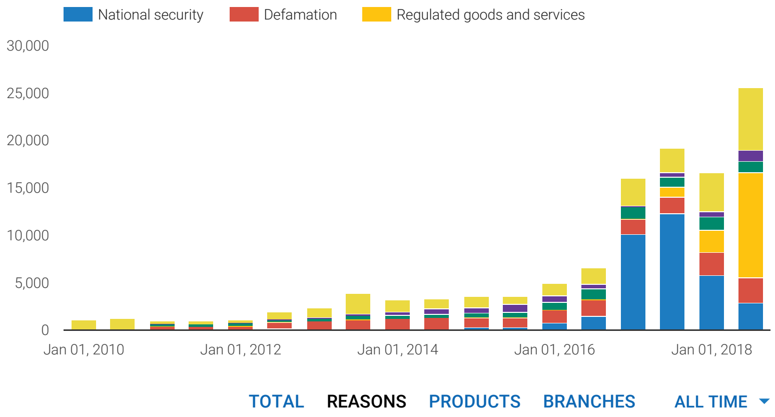
Please note that data will be stored in accordance with Google's privacy policy.

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

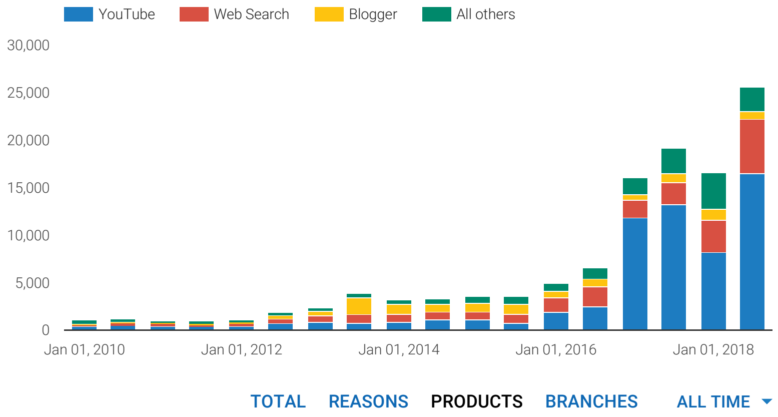
[252] YouTube, Hate speech policy.

[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 governments 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, 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 governments requests categorized based on reasons for content removal:

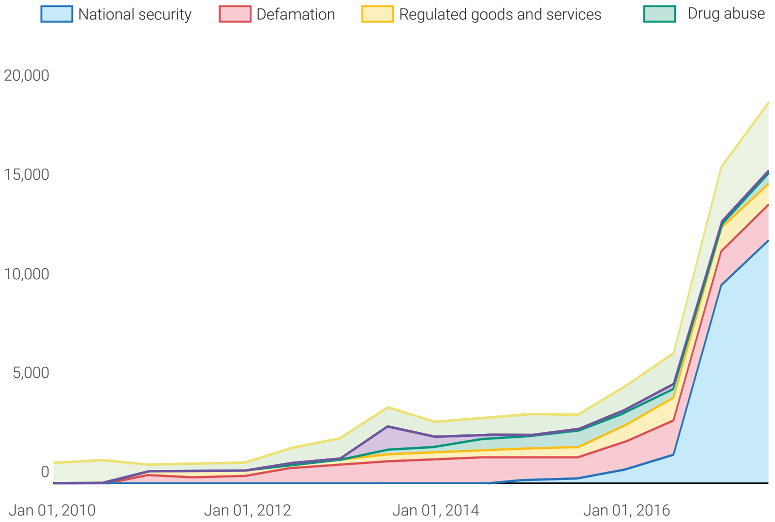


○ Google also displays the reasons for governments requests, categorized by products:



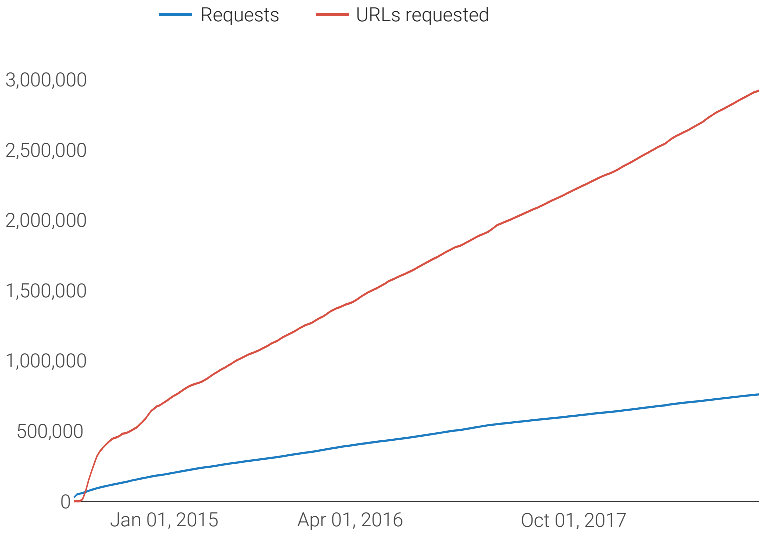
○ Google counts the reasons why governments ask for content removal. These data go back to December 2010:

Reasons cited for content removal

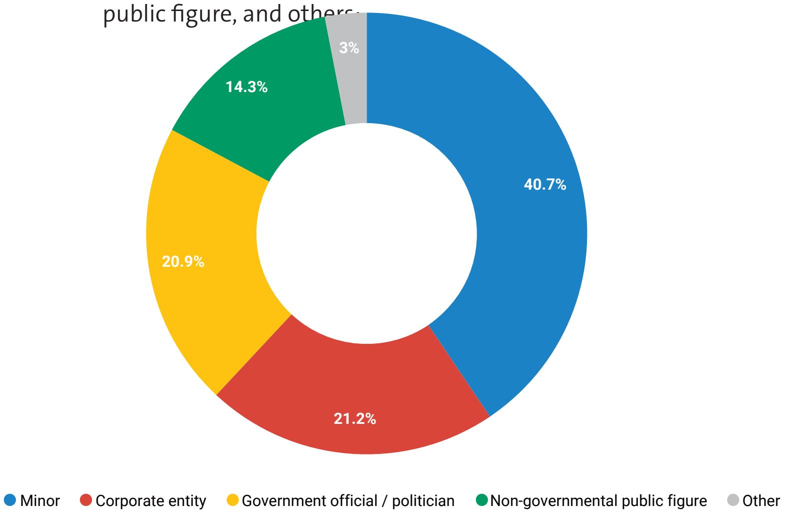


○ Google provides data on delist requests based on the European "right to be forgotten," the court-ordered right that allows users to ask search engines to remove certain results from queries. The search engine must comply if the links are "inadequate, irrelevant, or no longer relevant, or excessive. The search engine needs to take into account public interests 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 figure, 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 as 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:

[255] Microsoft Services Agreement (Published: March 1, 2018).

        - 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 and 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 https://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 Kingdoms.[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, since May 2014 and June 30, 2018, Microsoft received and processed 26,729 requests for 78,781 URLs. It accepted 32,725 requests (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 reserve 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 and states 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. Among others, the automation rules apply to automated abusive behavior, 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 which/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 mechanism:
        - 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 Twitter 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 individual who may be members. This targeting can happen in any manner (for example, @mentions, tagging a photo, and more).

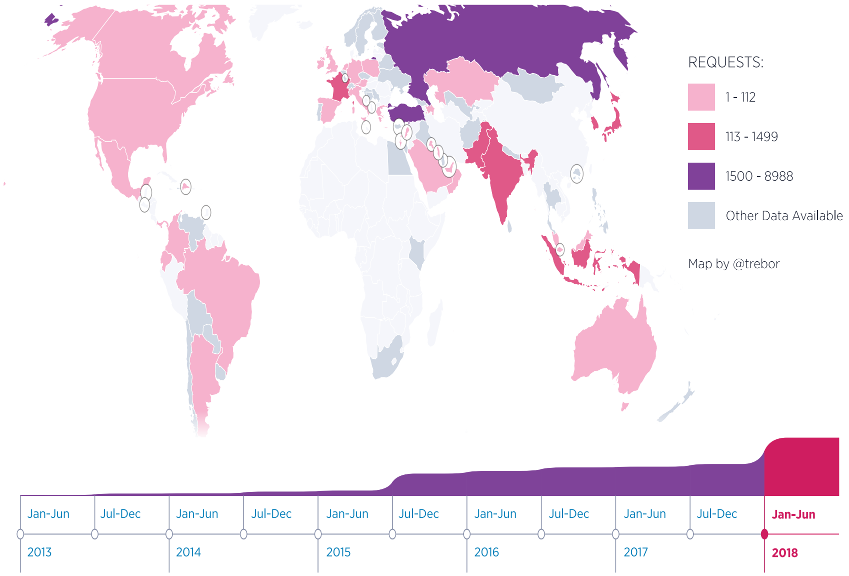
▪ The consequences of violating the rules vary depending on the severity of the violation and the person's previous record of violations. Twitter may ask users to remove an offending Tweet before they can Tweet again. Twitter may also suspend an account.

- 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 to 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;

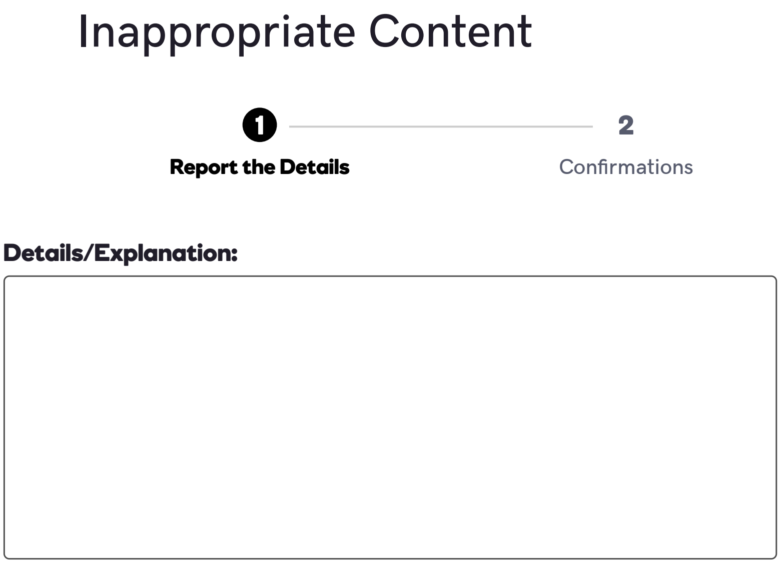
  ▪ Promotes, encourages or engages in terrorism, violence against people, animals, or property.

  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

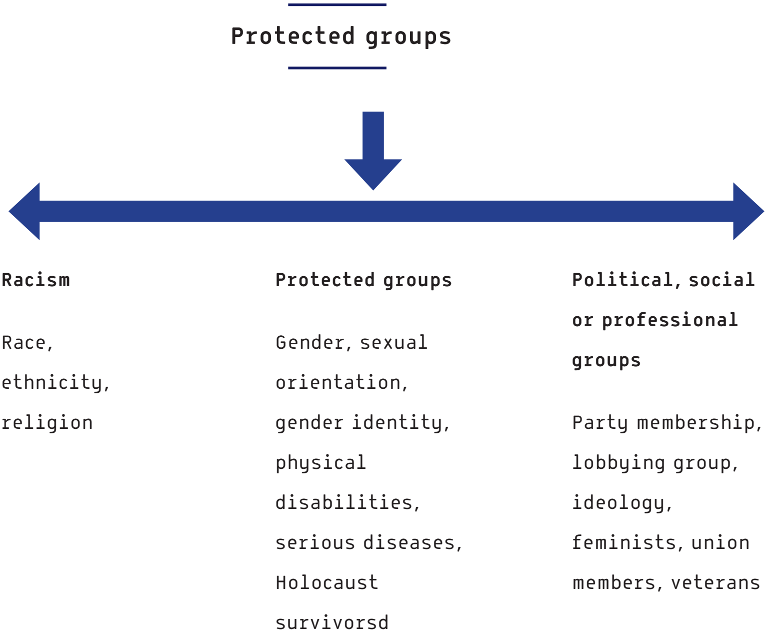
Applying the Proposed Model to Twitter's Community Standards

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. Speech that 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, Sept. 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."



Protected groups

Racism

Race, ethnicity, religion

Protected groups

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

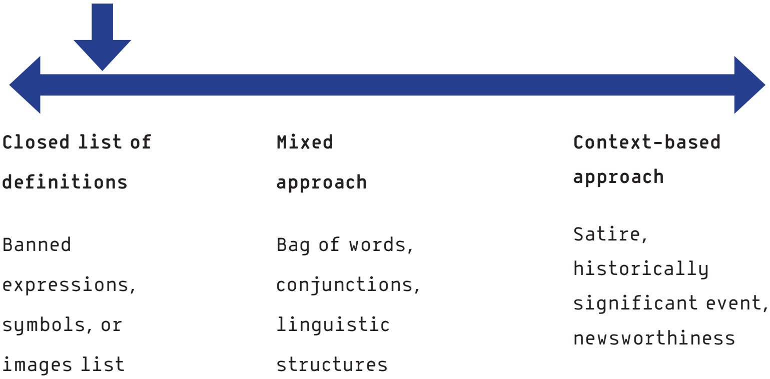
Political, social or professional groups

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

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 closed list of definitions.

[262] Twitter, *The Twitter Rules*.

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



Closed list of definitions

Banned expressions, symbols, or images list

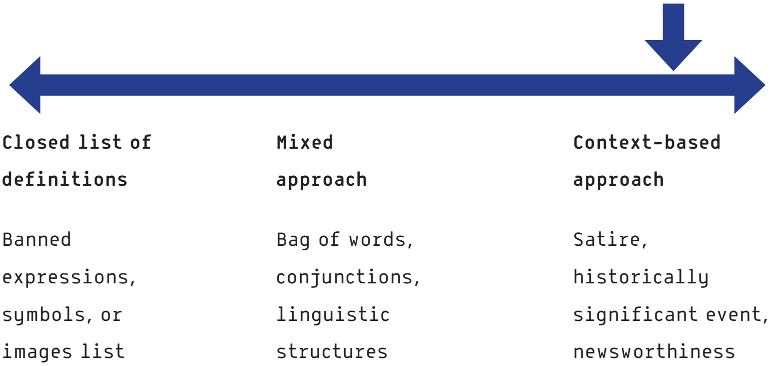
Mixed approach

Bag of words, conjunctions, linguistic structures

Context-based approach

Satire, historically significant event, newsworthiness

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]



Closed list of definitions

Banned expressions, symbols, or images list

Mixed approach

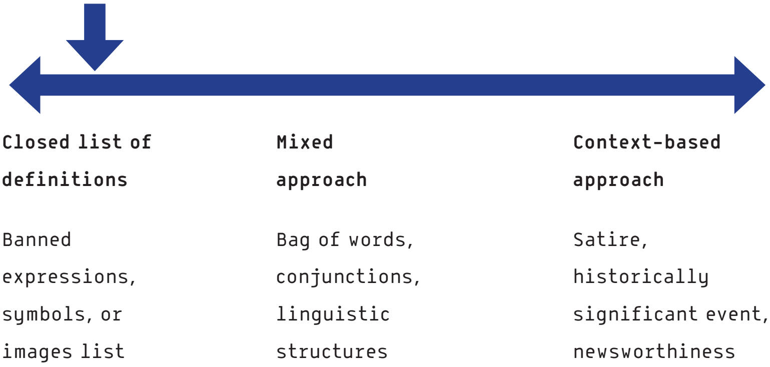
Bag of words, conjunctions, linguistic structures

Context-based approach

Satire, historically significant event, newsworthiness

[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]



Closed list of definitions

Banned expressions, symbols, or images list

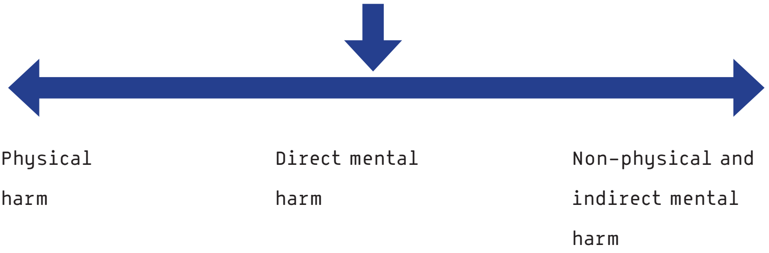
Mixed approach

Bag of words, conjunctions, linguistic structures

Context-based approach

Satire, historically significant event, newsworthiness

3. **The Speech could cause harm to an individual:** To ensure that users feel safe to express 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."



Physical harm

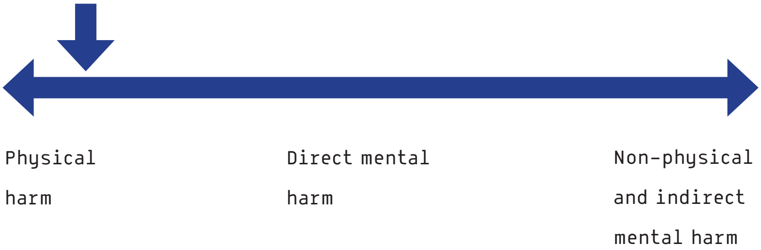
Direct mental harm

Non-physical and indirect 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]



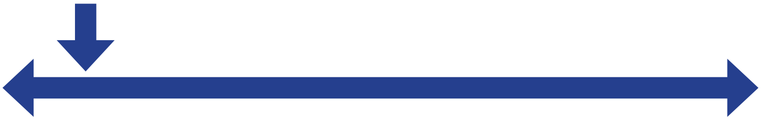
4. **The speaker intends to harm**

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 does 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*.

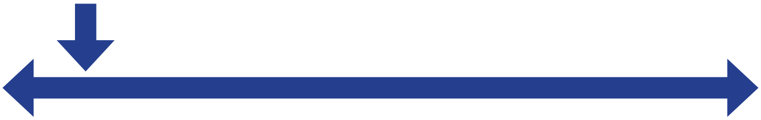
[267] According to Sellars, *supra* note 27, Twitter used to address underlying intent in its policies against conduct that promotes violence or directly attacks a group.

[268] Twitter, *Violent Threats and Glorification of Violence,* https://help.twitter.com/en/rules-and-policies/violent-threats-glorification (last visited: March 27, 2019).



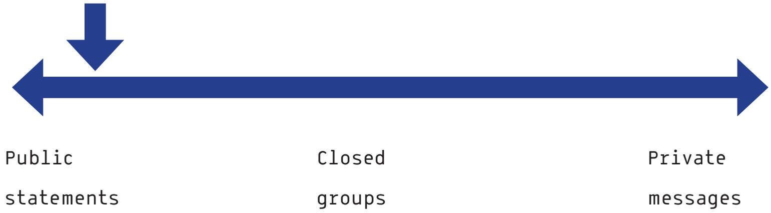
5. **The speech incites to socially undesirable action**

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 public nature of the speech**: 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.

Adv. Rotem Medzini, a former researcher at the Israel Democracy Institute, is 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. His research deals with internet policy, data protection, privacy, and content regulation.

Dr. Tehilla Shwartz Altshuler is a senior fellow at the Israel Democracy Institute and head of the Institute's 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.

This policy paper is the product of a collaboration between the Israel Democracy Institute and Yad Vashem the World Holocaust Remembrance Center. Its goal is to define practical co-regulatory tools for dealing with hate speech on social media platforms, content-wise and procedure-wise while balancing between the need for protection and defense of freedom of expression and the moral obligation to fight hate speech against various populations.

www.en.idi.org.il

0 4500001186 1
450-1186 דאנאקוד

ISBN:
978-965-519-257-5

June 2019