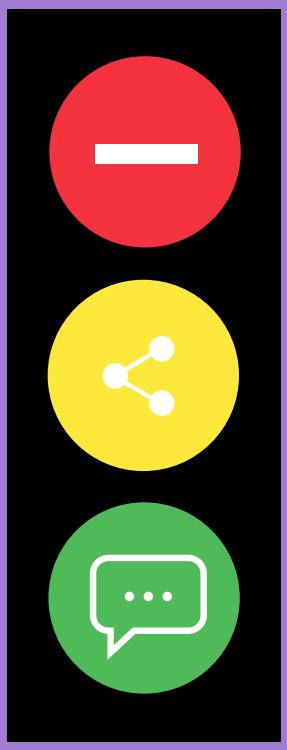
Proposals for Improved Regulation of Harmful Online Content



SUSAN BENESCH
Faculty Associate, Berkman Klein Center for Internet & Society
Executive Director, Dangerous Speech Project

INTRODUCTION

In its early life the internet inspired optimism that it would improve the world and its people, but that has been supplanted by alarm about harmful, often viral words and images. Though the vast majority of online content is still innocuous or beneficial, the internet is also polluted by hatred: some individuals and groups suffer harassment or attacks,[1] while others are exposed to content that inspires them to hate or fear other people, or even to commit mass murder.[2]

Hateful and harmful messages are so widespread online that the problem is not specific to any culture or country, nor can such content be easily classified under terms like "hate speech" or "extremism": it is too varied. Even the people who produce harmful content, and their motivations for doing so, are diverse. Online service providers (OSPs) [3] have built systems to diminish harmful content, but those are inadequate for the complex task at hand and have fundamental flaws that cannot be solved by tweaking the rules, as the companies have been doing so far. The stakeholders who have the least say in how speech is regulated are precisely those who are subject to that regulation: internet users.[4] "I've come to believe that we shouldn't make so many important decisions about speech on our own," Mark Zuckerberg, the CEO and a founder of Facebook, wrote last year.[5] He is correct.

Daunting though the problem is, there are many opportunities for improvement, but they have

---

1. *See,* e.g. European Commission, Directorate-General for Communication, Special Eurobarometer 452, Media Pluralism and Democracy (November 2016), at 17, https://perma.cc/PWL9-A6JV, reporting that "A large majority of those who follow or participate in debates has heard, read, seen or themselves experienced cases where abuse, hate speech or threats are directed at journalists/bloggers/people active on social media (75%)"; see also National Society for the Prevention of Cruelty to Children, Online abuse: facts and statistics, https://web.archive.org/web/20180401205802/https://www.nspcc.org.uk/preventing-abuse/child-abuse-and-neglect/online-abuse/facts-statistics/ (last visited February 15, 2018); Maeve Duggan, *Online Harassment 2017*, Pew Research Center July 11, 2017, https://perma.cc/EV4Q-ZLT9 (reporting a survey in which 62% of U.S. respondents regarded online harassment as a major problem and 40% had experienced it themselves); Steve Stecklow, *Why Facebook is losing the war on hate speech in Myanmar*, Reuters, Aug. 15, 2018, https://perma.cc/3CK3-4PA9; United Nations, Human Rights Council, *Detailed findings of the Independent International Fact-Finding Mission on Myanmar*, A/HRC/42/CRP.5, September, 16 2019, https://perma.cc/MJH7-K23Z
2. Jacob Asland Ravndal, *The Online Life of a Modern Terrorist: Anders Behring Breivik's Use of the Internet*, VOX Pol, Oct. 24, 2014, https://perma.cc/FTG4-URTC; Jessica Schulberg, Luke O'Brien, and Oliver Mushtare, *The Neo-Nazi Podcaster Next Door,* HuffPost, Feb. 7, 2019, https://perma.cc/4DBJ-CFB7; Adam Taylor, *New Zealand Suspect Allegedly Claimed 'Brief Contact' with Norwegian Mass Murderer Anders Breivik,* Washington Post, Mar. 15, 2019, https://perma.cc/P9ED-JE6C.
3. In this paper, "online service providers" (OSPs) refers to companies that host and disseminate user-generated content online and attempt to limit harmful content. Google, Facebook, and Twitter are the best known and most discussed (at least in the United States), but there are many others, large and small, including Reddit, Automattic, Bytedance, and companies that build and maintain chat apps, niche social media platforms, or online games.
4. Rebecca MacKinnon developed this idea in a 2012 book, and others have since joined her in calling for some kind of oversight of companies' governance of the speech of billions. *See e.g.* Rebecca MacKinnon, The Consent of the Networked (2012), Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, Harvard Law Review 131 (2017); Tarleton Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (2018)
5. Mark Zuckerberg, *The Internet Needs New Rules. Let's Start in These Four Areas,* Washington Post, Mar. 30, 2019, *available at* https://washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html.

been largely overlooked. The widespread distress about it is itself an opportunity, since that means millions of people are paying attention, and it will take broad participation to build online norms against harmful content. Such mass participation is neither far-fetched nor unfamiliar: many beneficial campaigns and social movements have been born and developed thanks to mass participation online.[6]

This paper offers a set of specific proposals for better describing harmful content online and for reducing the damage it causes, while protecting freedom of expression. The ideas are mainly meant for OSPs since they regulate the vast majority of online content; taken together they operate the largest system of censorship the world has ever known, controlling more human communication than any government.[7] Governments, for their part, have tried to berate or force the companies into changing their policies, with limited and often repressive results.[8] For these reasons, this paper focuses on what OSPs should do to diminish harmful content online.

The proposals focus on the rules that form the basis of each regulation system,[9] as well as on other crucial steps in the regulatory process, such as communicating rules to platform users, giving multiple stakeholders a role in regulation, and enforcement of the rules.

6. Zeynep Tufecki describes many of these in a 2018 book, though she also points out that the relative ease and speed of mass organizing online can make it harder to sustain social movements. *See* Zeynep Tufecki, Twitter and Tear Gas, The Power and Fragility of Networked Protest (2018).
7. The only government whose censorship system could rival the companies' in number of users or volume of content regulated is that of China, which has fewer than one billion people online *see e.g.* Jon Russell, *China reaches 800 million internet users*, TechCrunch, October 21, 2018, https://perma.cc/WZ98-PSZN. Facebook alone has more than 2.3 billion regular monthly users *see* Facebook, https://perma.cc/5FHJ-QJ8W. YouTube has nearly two billion users, and sees more than 400 hours of video posted during every minute *see* Danielle Abril, *YouTube Nears Major Milestone Amid Emphasis on Subscriptions,* Fortune, February 4 2019, https://perma.cc/U56H-77EX; Google Inc., *Monetization systems or 'the algorithm' explained,* Youtube Help, https://perma.cc/NH4E-533S.
8. William Echikson and Olivia Knodt, Germany's NetzDG: A Key Test for Combating Online Hate (2018), https://perma.cc/L2XQ-2A7U; Anthony Cuthbertson, *Pakistan Lifts Three-Year YouTube Ban on the Condition Censors Can Request Content Removal,* Newsweek, Jan. 19, 2016, https://perma.cc/MPV2-KB2W
9. I use the terms "moderate" and "regulate" to refer to the OSPs' myriad decisions to delete or keep content on their platforms, following Kate Klonick's wise practice. "Regulate" is not limited to government action here, especially since, as Klonick argues, OSPs now govern (or regulate). Kate Klonick, *supra* note 5, at 1601.

PART I. SUBSTANTIVE STANDARDS

To regulate behavior for collective benefit and to diminish social damage that it causes, it is best to define the behavior(s) in question clearly and to identify the harm that regulation is intended to prevent. OSPs have done significant work to diminish harmful content recently, responding to pressure from governments and the public.

Their moderation systems are deeply flawed, however. Rules are imprecise[10] and inconsistently enforced.[11] Enforcement is largely limited to two reactive methods—removing content and removing accounts—which constitute a blunt instrument that has little chance of achieving durable improvement by means of behavior change, i.e. diminishing the rate at which new harmful content is posted. Removing or 'taking down' content in industry parlance, is a necessary and important tool for content moderation, but is insufficient on its own.

Finally, most of the companies govern largely in secret. They make and implement their rules with only scant input from the people whose self-expression and access to information they restrict.[12]

A. Identifying forms of harmful content and harms

There are many forms of damaging content online, and they inflict almost as many types of harm, from causing emotional distress to inspiring mass murder. To be effective, regulation of harmful online content must therefore be both clear and complex. The following list of types of harmful online content gives a sense of its variety:[13]

- "Hate speech"
- Celebration of terrorist acts or violence
- Content designed to recruit extremists or terrorists

10. David Kaye, Speech Police: The Global Struggle to Govern the Internet (2019) about the Twitter Rules, "It's a vast and open-ended set of proscriptions."
11. David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, United Nations, Apr. 6, 2018, at 10, https://perma.cc/R5F3-YXSD.
12. Two significant exceptions to this are Reddit and the Wikimedia Foundation, which use what Robyn Caplan calls "the community-reliant approach": the company sets some high-level rules as a baseline, but relies on volunteers (who vastly outnumber those companies' employees) to both enforce its rules and establish additional norms and guidelines for various segments of the sites. *See* Robyn Caplan, Content or Context Moderation?, https://perma.cc/5FS3-4F2Z. And in May 2020, the video game streaming platform Twitch established a new advisory council, half of whose members are active streamers on Twitch. *See* Adam Smith, *Twitch Launches Safety Advisory Council to Help Clean Up Its Platform*, The Independent, May 15, 2020, https://perma.cc/HTU6-UHV2..
13. Scholars and researchers have developed several taxonomies of harmful online content, and OSPs' publicly available rules list types of harmful content in order to prohibit them, though most companies maintain more detailed taxonomies for internal use. For some examples, *see, e.g.,* Women's Media Center, *Online Abuse 101,* https://perma.cc/H5A8-5C82; Internet and Jurisdiction Policy Network, Content and Jurisdiction Program Operational Approaches (2019) at 20-26, https://perma.cc/ACR2-FC65; Facebook, *Community Standards,* https://perma.cc/ZC6T-KMEJ.

- Content to organize extremists or terrorists
- Credible threats of violence
- Graphic depictions of violence
- Fake accounts/impersonation
- Incitement to violence
- Instructions for making or using weapons of mass violence
- Dangerous speech[14]
- Bullying
- Harassment
- Abetting/promoting self-harm or suicide
- Sexual exploitation of children
- Nonconsensual or unsolicited pornography
- Defamation
- Doxing[15]
- Disinformation and deepfakes[16]
- Incitement to hatred of an identity group, which often includes falsehoods

It can be difficult to classify content into even these relatively granular categories, for several reasons. First, some of the categories (like the last two) overlap. Also, some content is not exclusively harmful: its presence online may also be constructive or beneficial. For example, human rights activists post video recordings of graphic police violence to denounce such conduct, in the hope of diminishing it,[17] and law enforcement agencies gather useful intelligence from some terrorist content.[18] Also, some painful content has historic or artistic value, such as the famous 1972 photograph by Nick Ut of Phan Thi Kim Phuc, a nine-year-old Vietnamese girl who was running naked while napalm from an airstrike burned into her back and side. When a Norwegian writer, Tom Egeland, posted it in 2016 as one of "seven photographs that changed the history of warfare," Facebook removed the image under the company's policy against nudity. That decision elicited protests by prominent Norwegian politicians, journalists, the Norwegian prime minister, Facebook users around the world, and Kim Phuc herself, who survived the burns and now lives in Canada. Finally conceding that the historical importance

14. "Dangerous speech," my own coinage, is any form of expression (speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group. For details including reasons why this category is useful, see dangerousspeech.org.
15. The term "doxing," derived from the word "documents" and its abbreviation "docs," means posting individuals' private information online, to expose them to harassment and attack by others.
16. Deepfakes are AI-generated videos or images that purport to show events or statements that never happened. They can be extremely difficult to identify as false. The word is a portmanteau of "deep learning" and "fake."
17. Jillian C. York, *Companies Must Be Accountable to All Users: The Story of Egyptian Activist Wael Abbas,* Electronic Frontier Foundation, Feb 13, 2018, https://perma.cc/6TJG-VHTF; David Uberti, *How Smartphone Video Changes Coverage of Police Abuse*, Columbia Journalism Review, April 9, 2015, https://perma.cc/9Y7P-E934.
18. But *see* Jessica Stern and J. M. Berger, ISIS: The State of Terror (2016), at 140. They argue that in most cases the intelligence to be gathered is not valuable enough to justify allowing terrorist content to remain online.

of the photograph outweighed the harm of depicting a naked child in this specific case, Facebook reversed its decision.[19]

PROPOSAL 1:

In order to prevent harm more effectively, companies should classify harmful online material not only by its content, but also by the harm it engenders. They should explain to users which forms of harm they seek to prevent, and to what degree.

As Facebook's decisions regarding the Kim Phuc photograph demonstrate, many key content moderation decisions are ultimately based (or should be based) not only on the content itself, but on the likely effects of its presence online; that is, on estimating harms and balancing them against possible benefits.

All systems of regulation, including bodies of law, are based on decisions about which harms should be suppressed and which can be tolerated. Companies' decisions on harm-balancing are consequential for billions of people who use their platforms. They should be both explicit and transparent, since people are more likely to follow rules whose purpose they understand.[20] Companies should explain which harms they have chosen not to tolerate, and why, and which content seems to produce those harms. Equally important, they should explain to users which harms they have chosen not to try to prevent, and why.

Only a few companies explain their moderation policies in terms of what damage they seek to forestall, and even then, in a limited way.

Facebook, for example, gives the following policy rationale: "We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence."[21] It does not mention other harms such as emotional distress, or decreased participation in civic life and discourse,[22] so users cannot know whether Facebook considers these tolerable, doesn't believe they are real, or simply chose not to mention them. Twitter says it bans what it calls "hateful conduct" (a narrower category than "hate speech," a term it does not use) because that content can curb the freedom of expression of those it denigrates; that is, it can "silence the voices of those who have been historically marginalized."[23] Twitter mentions no other

19. Sam Levin, Julia Carrie Wong, and Luke Harding, *Facebook Backs Down from 'Napalm Girl' Censorship and Reinstates Photo,* The Guardian, Sept. 9, 2016, https://perma.cc/J75H-AV94.
20. *See e.g.* M. E. Tankard & E. L. Paluck, *Norm Perception as a Vehicle for Social Change*, 10.1 Social Issues and Policy Review 181 (2016), https://doi.org/10.1111/sipr.12022.
21. Facebook, *supra* note 13, at sec. 12, "Hate Speech."
22. Jeremy Waldron, The Harm in Hate Speech (2014), at 5: "Hate speech is both a calculated affront to the dignity of vulnerable members of society and a calculated assault on the public good of inclusiveness."
23. Twitter Inc., *Hateful Conduct Policy,* Twitter Help Center, https://perma.cc/2PB6-62BH.

harm. YouTube gives no public rationale for its hate speech policy.[24]

Another reason to link policies with harms is that many forms of content inflict more than one type of harm, and often they can best be prevented with entirely different methods. For example, racist, anti-Semitic, or terrorist recruitment content can be deeply distressing to many people, and attractive or convincing to others. The former harm can be prevented by hiding the content—as users can do for themselves on some platforms, by means of filtering or blocking software. To prevent the latter harm, removal isn't sufficient on its own, since the same recruiting material can invariably be found somewhere else online. It is worth trying and testing other methods, such as pointing users who seem to be vulnerable to recruitment toward content designed to steer them away from hatred or extremism. The eponymous Redirect Method[25] is one such effort.

One more reason to classify content by the harms that it may cause is that not all harms should be eliminated, even if it were possible. A significant degree of offensiveness, for example, should be tolerated to protect freedom of expression, especially political speech.

It would be interesting to discover, also, whether every rule prohibiting a type of online content can be linked to a particular harm or harms that such content seems to engender among other users of a platform. It is possible that some rules are simply normative commitments by a company's leaders and not related to any harm. If so, this too should be made explicit.

As noted above, harms are nearly as varied as damaging content. Here are some examples:[26]

- exploitation of children
- mental or emotional distress (caused by content not related to the viewer)
- mental or emotional distress caused by a targeted or personal attack
- fear of being personally assaulted, due to a credible threat
- increased likelihood of self-harm
- violation of privacy
- damage to personal reputation
- economic harm to individuals or groups (e.g., job loss)
- silencing (decreased participation in online discourse)
- diminished participation in civic and public life[27]
- increased tendency to hate or fear, discriminate against, or endorse violence against other people

24. Google, *Hate Speech Policy,* YouTube Help, https://perma.cc/3Q4M-YDC6.
25. *Redirect Method*, https://redirectmethod.org, *archived at* https://perma.cc/A4QY-Q4TC See also Lydia Dishman, *Google Algorithms and Human Psychology: How Jigsaw Rescues Teens from ISIS Recruiters*" Fast Company, Jan. 28, 2019, https://perma.cc/26L3-DCQ3.
26. For another taxonomy of harm caused by online content, *see, e.g.*, Women's Media Center *supra* note 13; *see also* Maeve Duggan, *Online Harassment 2017*, Pew Research Center, July 11, 2017, https://perma.cc/9F6H-DJM7.
27. Waldron, *supra* note 22.

- deterioration of the tone of online discourse
- normalizing violence and other harmful offline behavior
- convincing people of falsehoods
- collective social or national harms enumerated in Article 19(3)(b) of the International Covenant on Civil and Political Rights: damage to national security, public order, public health, or morals

Finally, to make their efforts to diminish harms more effective, companies should consider classifying harms by severity or gravity. This would allow them to build triage systems and to focus on responding first, or most quickly, to the worst examples.

B. "Hate speech"

The term most often used by the public, government officials, and academics to describe harmful online content is "hate speech." In spite of its wide use there is no consensus—in law, OSP rules, or colloquial use[28]— about what falls into that category, except egregious examples. For many of those, moreover, the term "hate speech" is not necessary, since they also constitute other speech acts (such as incitement to violence) that are similarly defined in multiple bodies of law. As Andrew Sellars observed in a paper, "Defining Hate Speech," in which he offers important and useful ideas toward a definition (but doesn't quite propose one), "surprisingly little work appears to have been done to define the term 'hate speech' itself. Without a clear definition, how will scholars, analysts, and regulators know what speech should be targeted?"[29] Confusion over which speech to include has led to many cases of mistaken removal, and failure to remove hateful content.[30]

PROPOSAL 2:

OSPs should clearly define which content they regulate, describe boundaries between what is prohibited and what is permitted, and explain how they take context into account.

To contribute to a clearer and more uniform definition of online hate speech, this section summarizes existing OSP rules, national laws, and international human rights law. Each of these has special

28. *See* interview with Kenan Malik, in The Content and Context of Hate Speech: Rethinking Regulation and Responses (Michael E. Herz and Péter Molnár, eds., 2012), 81: "If you look at hate speech laws across the world, there is no consistency about what constitutes hate speech."

29 . Andrew F. Sellars, *Defining Hate Speech* (Dec. 1, 2016), Berkman Klein Center Research Publication No. 2016-20, 4, https://perma.cc/T6UT-88TN. For more detail on the variety of definitions, *see* Susan Benesch, *Defining and Diminishing Hate Speech*, in State of the World's Minorities and Indigenous Peoples 2014, (Minority Rights Group, 2014), *at* 18, https://perma.cc/H7GD-EA2P. As pointed out by ARTICLE 19, a human rights NGO, there is no consensus definition of the term. See ARTICLE 19, Hate Speech Explained: A Toolkit (2015), https://perma.cc/XB2Y-UTZ6.

30. Davey Alba, *Defining 'Hate Speech' Online is an Imperfect Act*, Wired, Aug. 22, 2017, https://perma.cc/F4T9-563S.

relevance: platforms' own rules form the basis of most moderation now underway; companies are obliged to comply with national laws wherever they operate; and international human rights law could serve as a universal basis for content moderation, as David Kaye, the U.N. Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, has proposed.[31]

The human rights organization ARTICLE 19 notes that "'[h]ate speech' is an emotive concept which has no universally accepted definition in international human rights law."[32] Perhaps it is the emotive nature of "hate speech" that has helped make the term so popular, despite its ambiguity. It is often used to signal the reader's (or listener's) outrage, as much as the author's intent. As the writer and human rights advocate Salil Tripathi put it, "[f]rom speech that promotes hatred, hate speech has come to mean speech you hate. A nebulous term whose meaning varies from person to person, 'hate speech' is increasingly being used to vilify words and speech that we disagree with, and hence hate, expanding its meaning significantly from what it was meant to be—speech that encourages people to hate others."[33] Even "hate" itself is somewhat ill-defined, as the legal scholar Robert Post has pointed out,[34] a and it is not clear whether the "hate" in "hate speech" refers to the state of mind of the speaker/author, to the likely increase in hateful thoughts among a receptive audience, to the speech's capacity to make people (those it attacks or purports to describe) feel hated—or, as Tripathi argues, to an expression of outrage or disagreement with the speech. The terms "hate" or "hatred," where they are defined in law at all, are usually understood narrowly. For instance, Canada's criminal code provision against the "willful promotion of hatred" must be "construed as encompassing only the most severe and deeply felt form of opprobrium," the Canadian Supreme Court found in the landmark case of James Keegstra, a public school teacher who told his students that Jews are an evil people who "created the Holocaust to gain sympathy."[35]

Many would simply say, as U.S. Supreme Court Justice Potter Stewart famously wrote about pornography, "I know it when I see it."[36] But that would not provide consensus on what hate speech is, for many reasons. First, people identify it differently, according to their cultural backgrounds and normative commitments. Second, the meaning—and the dangerousness or capacity to bring about harm—of almost any putative hate speech depends on the context in which it is expressed or disseminated.[37] Third, people can be maddeningly inventive in expressing or fomenting hatred: often hateful content contains no slurs or telltale words, in part to evade detection,[38] but is still clearly understood by its intended audience and can be at least as vicious and powerful as content that contains obviously hateful language. In fact, sometimes coded language or images serve

31. Kaye, *supra* note 11.
32. ARTICLE 19, Self-Regulation and 'Hate Speech' on Social Media Platforms (2018), at 6, https://perma.cc/AVK8-9N7B.
33. Salil Tripathi, *Hate Speech*, Seminar 716 (April 2019), 24, https://perma.cc/P5DA-NHN2.
34. Robert Post, *Hate Speech*, in Extreme Speech and Democracy (Ivan Hare and James Weinstein, eds. 2009), at 123
35. R. v. Keegstra, [1990] 3 S.C.R. 697 (Can.) Part VII(D)(iii)(a) (Dickson, C.J.), https://perma.cc/Y5FT-YT96.
36. Jacobellis v. Ohio, 378 U.S. 184, at 197 (Stewart, J., concurring).
37. *See* Dangerous Speech Project, Dangerous Speech: A Practical Guide (2018) at 19, https://dangerousspeech.org/guide, *archived at* https://perma.cc/Z5SF-45WT.
38. Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths, *A Web of Hate: Tackling Hateful Speech in Online Social Spaces,* in Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety, 2016, https://perma.cc/5TV9-YK5W. *See also* Sellars, *supra* note 29, at 4: "When talking of hate speech, a shocking degree of the discussion—be it academic or in public discourse—looks solely to finding specific words or phrases that the observer believes signal the presence of hate speech. Is that a sound strategy?

as a kind of social glue, an in-joke that binds a group of people together. This is one reason why extremist and hate groups are heavy users of hand gestures with in-group meanings, or polysemic memes such as Pepe the Frog. Fourth, slurs are sometimes reclaimed by members of the group whom they ostensibly describe, who use them in non-offensive ways. Fifth, activists and targets of hate sometimes deliberately repeat it in order to denounce it or call it out, and that content is often mistakenly censored.[39] It is therefore difficult to write—and harder to apply—rules prohibiting and accurately classifying "hate speech," and even harder to detect it reliably with automated software tools (called classifiers or simply algorithms). This is vital to remember, since it is otherwise tempting to try to rely on software to detect and automatically remove "hate speech."

The slipperiness of the term can also pose a serious threat to freedom of expression, since it makes it easy for governments to use it to prosecute their political opponents or minority groups. In Hungary, for example, where hateful speech against Roma is all too common and has led to violent attacks on members of that group, Roma have been prosecuted for "anti-Hungarian hate speech."[40] In Kazakhstan, a law against inciting religious hatred has been used to imprison atheists, human rights activists, and Muslims, in one case for reading a publicly available book. "Ablaykhan Chalimbayev spent five years in a Kazakh prison for quoting a commentary on the Quran" under the law against religious hatred, as the Danish lawyer and human rights advocate Jacob Mchangama noted with concern.[41] The term "hate speech" can also be used as a political weapon, as it was during Kenya's 2013 presidential campaign when some Kenyans felt that it was used to suppress debate, just when it was more necessary than ever.[42]

There is a common thread in virtually all definitions of hate speech, which is that it denigrates or attacks people based on some kind of shared identity or membership in certain kinds of groups. Therefore no matter how emphatically a person declares, "I hate you!," that is not "hate speech" if there is no reference to a group.

Laws and definitions of hate speech usually list specific types of groups or shared identities, such as ethnicity, religion, race, or nationality/national origin. Categories such as gender, age, sexual orientation, immigration status, disease, and/or disability are included in some definitions but not others. This has led to heated debates over which categories should "count." Definitions vary also with regard to how severe a speech act must be to constitute "hate speech": inciting violence against a member or members of a group; dehumanizing them; suggesting that they are inferior, dangerous, or too numerous (and therefore threatening); or insulting them in another way.

---

39. Tarleton Gillespie, *supra* note 4.
40. Milkos Haraszti, *Foreword: Hate Speech and the Coming Death of the International Standard before It Was Born (Complaints of a Watchdog),* in The Content and Context of Hate Speech: Rethinking Regulation and Responses (Michael Herz and Peter Molnar, eds., 2012).
41. Jacob Mchangama, *The U.N. Hates Hate Speech More than it Loves Free Speech,* Foreign Policy, Feb. 28, 2019, https://perma.cc/7HMS-YQRL. *See also* Andrey Grishin, *How Kazakhstan's Anti-Extremism Blacklist Forces Activists, Bloggers and Opposition Politicians into the Shadows,"* OpenDemocracy, Aug. 7, 2018, https://perma.cc/N2B6-GS63.
42. Patrick Gathara, *The Monsters Under The House,* Gathara's World (blog), Mar. 10, 2013, https://perma.cc/3EBN-4GGP.

1. OSP rules on "hate speech"

OSPs define hate speech quite differently, and some choose not to ban it at all. Reddit CEO Steve Huffman explained why Reddit does not, saying "hate speech is difficult to define" and "it's impossible to enforce consistently."[43] Among those that do, YouTube calls it "promoting violence or hatred" and Facebook describes it as "violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation." As discussed above, Twitter doesn't ban "hate speech." Instead it prohibits a narrower category that it calls "hateful conduct," by which it means conduct that "promote[s] violence against or directly attack[s] or threaten[s] other people." Twitter also separately prohibits the use of hateful imagery or symbols in a profile or header image. At each company, the rules have evolved over time, and changes often come in response to public controversies over specific pieces of content.

Take for example Facebook's announcement, a few days after the March 2019 massacre at two mosques in Christchurch, New Zealand, that it would ban expressions of white nationalism and white separatism. Facebook had previously considered those legitimate speech, distinguishing them from white supremacy, which it did identify as hate.[44] The Christchurch killer live-streamed video of himself committing the massacre on Facebook, and the recording was posted on many sites online; he also posted a "manifesto" in which he repeated the white supremacist claim that Muslim immigrants pose an existential threat to "Europeans" like himself. Facebook's decision led commentators to wonder whether it would apply the same new criteria to other nationalists, not only white ones. As Salil Tripathi commented on Facebook's announcement, "the arbitrariness of social media companies in deciding what goes on air and what doesn't, is deeply troubling. [...] Would it do the same for Hindu nationalist/ Muslim fundamentalist pages?"[45]

Companies also include different identity groups in their definitions of "hate speech," effectively offering extra protection to certain groups but not others. Facebook and YouTube include caste, for example, and YouTube adds veteran status.[46] Finally, Facebook is unusual in describing three "tiers" of hate speech, all of which it ostensibly prohibits. The first is violent or dehumanizing speech, the second is speech claiming that a member or members of another group are inferior or deficient, and the third refers to calls to segregate or exclude. The current public rules of Facebook, YouTube, Twitter, and several other platforms regarding "hate speech" or hateful conduct are presented in an appendix to this paper, for reference and comparison.

It is vital to note that each set of public rules is only the tip of a much larger iceberg, since most companies have more than one set of rules: the publicly available ones such as Facebook's "Community Standards," and a much more detailed manual that moderators use to make decisions.

---

43. Shoshana Wodinsky, *Reddit CEO Says It's 'Impossible' to Consistently Enforce Hate Speech Rules,* The Verge, July 9, 2018, https://perma.cc/59GH-LL25.
44. Tony Romm and Elizabeth Dwoskin, *Facebook Says It Will Now Block White-Nationalist, White-Separatist Posts*, Washington Post, Mar. 27, 2019, https://perma.cc/55FF-XACZ; Facebook, *Standing Against Hate,* Facebook Newsroom (blog), March 27, 2019, https://perma.cc/5CMV-NQUZ.
45. Salil Tripathi, *@saliltripathi* (Twitter), Mar. 29, 2019, https://perma.cc/V9J8-H4VE.
46. Facebook, *supra* note 13, at sec. 12, "Hate Speech."; Twitter Inc., *Hateful Conduct Policy*, Twitter Help Center, https://perma.cc/G3S6-32J2; Google, *Hate Speech Policy,* YouTube Help, https://perma.cc/9XBF-WCAH.

The latter are kept secret,[47] which greatly limits the extent to which outsiders can understand and critique the companies' actual governance of "hate speech" and other content—the way they define such content in practice.

Since those manuals make granular distinctions between prohibited and permitted content, they should be made accessible to outsiders, including users who want to understand the details, not merely the general standards.[48] The moderators' manuals give detailed instructions for applying the standards to specific, real cases, the way regulations are used to interpret statutes. Companies have long resisted releasing their manuals, saying that this would allow bad actors to "game the system"—to find ways of remaining just barely on the permissible side of a rule, or, more generally, ways of posting vicious or harmful content while avoiding takedown. These justifications are not persuasive for two reasons. First, laws constantly draw lines between prohibited and permitted behavior, and a line is drawn in a particular way because behavior anywhere on the permitted side of the line is considered acceptable. If it is not, the line should be moved. Second, users who are determined to post harmful content and evade removal can extrapolate where the lines are, by testing the system with a variety of posts from a variety of accounts. This is commonly done, for example, by coordinated propagandists in Myanmar, according to Michael Lwin, co-founder and managing director of Koe Koe Tech, a Yangon-based IT firm.[49]

Finally, OSPs should explain how they account for varied social, cultural, and political context when they make takedown decisions. The same hateful remark or frightening rumor can have a dramatically different capacity to influence people (and even catalyze action) in different contexts. Platforms like Facebook claim to use only one set of moderation rules for the entire world (or the large proportion of it in which they operate). Surely the rules prescribe different decisions as context changes, however. This, too, should be explained for those who are governed not so much by the "Community Standards" as by their tangible application to millions of pieces of content.

2. National laws on "hate speech"

Most bodies of national law do not mention the term "hate speech" at all, much less define it. Instead, some refer to speech acts such as incitement and discrimination—or unique to Rwanda, the vaguely and broadly defined offense of "ethnic divisionism."[50] Other laws focus on a variety of harmful consequences of speech, including insult, offence, humiliation, and degradation. Laws also identify unlawful speech by the intent of the speaker, the likely effect of the speech, and whether the speech

47. Facebook published a more detailed version of its rules in 2018, but they were not nearly as extensive or granular as the ones used by moderators. Josh Constine, *Facebook Reveals 25 Pages of Takedown Rules for Hate Speech and More,* TechCrunch, Apr. 24, 2018, https://perma.cc/G2N6-KWX2. In a few cases, parts of internal manuals for moderators have been leaked. *See, e.g.,* Julia Angwin and Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children,* ProPublica, June 28, 2017, https://perma.cc/BUY3-YDNU; Nick Hopkins, *Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence,* The Guardian, May 21, 2017, https://perma.cc/CX33-LAKC.

48. *See* Klonick, *supra* note 4, 1631, distinguishing between standards and rules.

49. Interview with the author, 2020.

50. Immigration and Refugee Board of Canada, Rwanda: Legislation Governing Divisionism and its Impact on Political Parties, the Media, Civil Society and Individuals (2007). RWA102565.E, https://perma.cc/7NTH-ZVQ3.

calls for action of some kind.

U.S. national law famously does not criminalize "hate speech." In fact, it protects the right to produce almost every form of it, under the First Amendment to the U.S. Constitution.[51] Only a very small subset of what would be considered "hate speech" by some definitions is criminalized, under the standard developed by the U.S. Supreme Court in the 1969 case of Brandenburg v. Ohio. Speech can be criminal if it is "directed to inciting or producing imminent lawless action" and is also likely to successfully incite or produce such action.[52] In other words, only incitement to violence that is likely to succeed quickly is prohibited. Thus U.S. law protects "hate speech" more than any other body of law in the world—and has been highly influential in the development of OSPs' moderation systems, since it was a formative influence on the people who designed them. As Kate Klonick observed, "American lawyers trained and acculturated in American free speech norms and First Amendment law oversaw the development of company content moderation policy. Though they might not have 'directly imported First Amendment doctrine,' the normative background in free speech had a direct impact on how they structured their policies."[53]

Other bodies of national law criminalize large swaths of the same speech that the U.S. First Amendment protects. For example, §135a of the Norwegian penal code defines "hate speech" very broadly, in terms of both prohibited actions and protected identity groups. Hate speech is defined as "threatening or insulting anyone, or inciting hatred or persecution of or contempt for anyone because of his or her (a) skin color or national or ethnic origin, (b) religion or life stance, or (c) homosexuality, lifestyle or orientation."[54] South Africa's hate speech law is one of the most detailed and comprehensive, specifying groups and attributes that are not found in other countries' legislation, such as pregnancy, marital status, conscience, language, skin color, and "any other group where discrimination based on that other ground (i) causes or perpetuates systemic disadvantage; (ii) undermines human dignity; or (iii) adversely affects the equal enjoyment of a person's rights and freedoms in a serious manner that is comparable to discrimination […]." [55]

Bhikhu Parekh has illustrated the diversity of national laws with a set of examples:

> "Britain bans abusive, insulting, and threatening speech. Denmark and Canada prohibit speech that is insulting and degrading; and India and Israel ban speech that incites racial and religious hatred and is likely to stir up hostility between groups. In the Netherlands, it is a criminal offence to express publicly views insulting to groups of persons. Australia prohibits speech that offends, insults, humiliates, or intimidates individuals or groups, and some of its states have laws banning racial vilification. Germany goes further, banning speech that violates the dignity of an individual,

51. U.S. Const. amend. I. "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the government for a redress of grievances."
52. *Brandenburg v. Ohio*, 395 US 444 (1969), https://perma.cc/82ET-LWBD.
53. Klonick, *supra* note 4.
54. The General Civil Penal Code (Act No. 10 of May 22, 1902, as last amended by Act No. 131, Dec. 21, 2005), University of Oslo Law Library Translated Norwegian Legislation online database, https://perma.cc/V6M5-MC4E.
55. Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000, c. 1, https://perma.cc/YQX5-DU8T.

implies that he or she is an inferior being, or maliciously degrades or defames a group."[56]

Germany also prohibits denying the Holocaust in a manner that could disturb the public peace[57] and prohibits disturbing "the public peace in a manner that violates the dignity of the victims by approving of, glorifying, or justifying National Socialist rule of arbitrary force."[58]

In sum, national laws on hate speech and related content vary greatly. Many of them are vague or broad enough to be difficult for OSPs to interpret, and to be subject to easy misuse by governments.[59]

3. International human rights law on speech

In a 2018 report to the UN Secretary General, Special Rapporteur on Freedom of Opinion and Expression David Kaye proposed that international human rights law serve as uniform guidelines for national laws on online content moderation. Private companies' rules have created "unstable, unpredictable, and unsafe environments," Kaye wrote. Human rights standards could be improved by the provision of "a framework for holding both States and companies accountable to users across national borders."[60] ARTICLE 19, an international freedom of expression organization, has made the same recommendation, arguing like Kaye that this would lead to more clear and consistent rules, more transparency about what the rules are and how they are applied, and more opportunity for oversight. an international freedom of expression organization, has made the same recommendation, arguing like Kaye that this would lead to more clear and consistent rules, more transparency about what the rules are and how they are applied, and more opportunity for oversight.[61]

I agree, with a caveat: that human rights law on speech is confusing and not always applicable to private companies. If properly interpreted and explained by experts, however, it could serve as an important source of standards for content moderation by companies.

56. Bikhu Parekh, Is There a Case for Banning Hate Speech, in THE CONTENT AND CONTEXT OF HATE SPEECH, (Michael Herz and Peter Molnar, eds., 2012), 37. Britain's prohibition on "insulting" speech was criticized for being too broad (especially after it was used for dubious prosecutions such as one of a university student for insulting a policeman's horse) and was removed by the Crime and Courts Act 2013.
57. German Criminal Code, Section 130(3): "Whosoever publicly or in a meeting approves of, denies or downplays an act committed under the rule of National Socialism of the kind indicated in section 6 (1) of the Code of International Criminal Law, in a manner capable of disturbing the public peace shall be liable to imprisonment not exceeding five years or a fine." https://web.archive.org/web/20200113180149/https://gesetze-im-internet.de/englisch_stgb/englisch_stgb.html
58. Id., Section 130(4).
59. Kaye, *supra* note 11, at 9. "The commitment to legal compliance can be complicated when relevant State law is vague, subject to varying interpretations or inconsistent with human rights law."
60. Id., at 14.
61. ARTICLE 19, *supra* note 32.

PROPOSAL 3:

International human rights law on speech should serve as a source of general standards for moderation of "hate speech" and other harmful content by companies—after it has been analyzed and interpreted for this purpose by outside experts.

Such interpretation is particularly needed regarding "hate speech" since that term is nearly absent from international law[62] and is not mentioned in the applicable core treaties and declarations, which refer instead to offensive, inciting, or discriminatory speech. Article 7 of the Universal Declaration of Human Rights states that all persons are entitled to protection against discrimination in violation of the Declaration—and against "any incitement to such discrimination."[63]

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) confers the right to freedom of expression and opinion. It also establishes that a state may prohibit expression only if the prohibition is: (1) provided by law, (2) necessary in a democratic society, and (3) in pursuit of one of the following aims: respect of the rights or reputations of others; or the protection of national security, public order, or public health or morals.[64] For this provision to be applied to OSPs, their rules must be understood as law. Indeed, Kaye and other scholars refer to the companies' own rules as "platform law,"[65] since they are used for governance.[66] How should the rest of the terms be understood and used by companies? Should they make decisions about the national security of countries around the world? Societies' public order? Morals? If so, shouldn't they consult with stakeholders in the relevant countries? Which ones, then, and on what terms? These questions need to be answered before international human rights law can offer standards for content moderation by companies, other than in vague and general terms.[67]

After Article 19 sets out the circumstances in which governments may prohibit expression, Article 20 sketches the types of expression that they must prohibit: "Any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."[68] This provision is unclear, in no small part because the distinctions between advocacy and incitement

62. Hate speech makes an appearance in international criminal law as a form of persecution, which, when sufficiently widespread and systematic, can constitute a crime against humanity.
63. Universal Declaration of Human Rights, Dec. 10, 1948, G.A. Res 217 A(III), U.S. Doc A/810 at 71 (1948).
64. International Covenant on Civil and Political Rights, G.A. Res. 2200A(XXI), 999 U.N.T.S. 171 (Dec. 16, 1966).
65. Orly Lobel, *The Law of the Platform,* 101 Minnesota Law Review 86 (2016), cited in Kaye, *supra* note 11.
66. Klonick, *supra* note 4; Gillespie, *supra* note 4.
67. Evelyn Aswad has explained how much of Article 19 of the ICCPR could be used by OSPs. See Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 Duke Law & Technology Review 26-70 (2018), https://perma.cc/9MHK-QG7Q. I am writing a forthcoming article, offering more ideas. *See also* ARTICLE 19, *Side-stepping Rights: Regulating Speech by Contract* (2018), at https://perma.cc/N6CH-FLF8; United Nations, General Assembly, Promotion and protection of the right to freedom of opinion and expression: note by the Secretary-General, A/74/486, October 9, 2019, https://perma.cc/3L4H-YC2W.
68. International Covenant on Civil and Political Rights, *supra* note 64

on the one hand, and between hatred and hostility on the other, are unclear, within and among bodies of law. Jacob Mchangama has described how the odd and confusing formulation of Article 20 emerged from its contentious drafting.[69]

Article 20 has been incorporated into bodies of national law only partially, or not at all. In light of the confusion it engenders, in 2011 and 2012 the United Nations High Commissioner for Human Rights oversaw an effort to clarify it.[70] This led to the Rabat Plan of Action,[71] which proposes a six-part threshold test for unlawful incitement. Because the test consists of factors that are often very difficult to determine online, such as the speaker's intent, it may be of limited applicability to content moderation.

Another core international human rights treaty is directly relevant to "hate speech," although it omits the term. The International Convention on the Elimination of all Forms of Racial Discrimination (ICERD)[72] calls on its parties to "declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin."[73] This is evidently a lower threshold than the ICCPR's, and would require restricting much more speech.

Moreover, such treaties set standards that are quite general, whereas content moderation requires highly specific, granular rules. This is especially true as moderation is conducted on a large scale and OSPs need to train thousands of moderators to make consistent decisions. If international human rights standards come to guide online content moderation, different platforms may derive quite different rules from them. How wide should the range of variability be?

Such questions should be resolved by experts, perhaps organized as an international council that would interpret international human rights law as it applies to content moderation by private companies. This group might be convened by the relevant UN special rapporteurs. Once international human rights law is explicated for use in private online content moderation, it can provide a useful set of universal standards.

Councils of outside advisors should not be composed only of human rights lawyers. For their recommendations to be feasible and realistic, they should include people with significant knowledge of how social media and other platforms work from the technical point of view, such as engineers, designers, and user experience (UX) researchers. In other words, the analysis of human rights law is only one area in which OSPs should seek and rely on guidance from non-government outsiders, including the people governed by their rules—their users.

---

69. Jacob Mchangama, *The Sordid Origin of Hate Speech Laws,* Policy Review, Dec. 1, 2011, https://perma.cc/T4NB-79A6.
70. United Nations, "Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement to Discrimination, Hostility or Violence," https://perma.cc/3HKH-PWUA.
71. "Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement to Discrimination" (United Nations, January 11, 2013), https://perma.cc/WHD6-U2G6.
72. UN General Assembly Resolution 2106A(XX), 21 December 1965, https://perma.cc/NY2D-7DEE.
73. UN General Assembly, *International Convention on the Elimination of All Forms of Racial Discrimination*, 21 December 1965, United Nations, Treaty Series, vol. 660, https://perma.cc/2EEV-NYWH.

PART II. PROCEDURAL STANDARDS

PROPOSAL 4:

OSPs should develop councils of non-government outsiders to review and advise them on their content moderation rules, on both broad (national or international) and local, granular levels. In this way, their detailed rules can be properly adapted to cultural contexts, as long as this "margin of appreciation" does not lead to violations of international human rights law standards.

There is a growing consensus, now even including Mark Zuckerberg,[74] that OSPs should not write and apply rules entirely on their own. By doing so, they have operated without an external check on their rules and deprived users of agency in the basis of governance. That produces, as David Kaye puts it, a "democratic deficit."[75] "The companies, as private stewards of public space," he writes, "interfere with the idea that their users are engaging in democratic culture. Users become subjects. In that sense, platform 'life' diminishes democratic culture even as it expands the possibilities of communication."[76] Without knowledge of the rules governing online spaces, and without any sense of representation in the making of those rules, people are less likely to obey rules against hate speech and other forms of harmful behavior.

Independent external review and oversight of OSPs' rules could well lead to better, more consistent regulation of harmful content online. This will require some bold experiments. First of all, many questions present themselves, such as who exactly should contribute to the rulemaking and rule-enforcing processes, how those people will be chosen, how much authority they will have, and how they will be held accountable.

Until recently, most OSPs formed only limited advisory bodies[77] such as Twitter's Trust and Safety Council which includes online safety and anti-hatred advocates and some researchers, including my organization. The council has no decision-making power at all; its members simply give intermittent advice at the request of Twitter staff, and sometimes the council learns of major policy changes only when Twitter announces them publicly.[78] Facebook has a Safety Advisory Board that includes some of the same members and plays a similar role, again without authority.[79]

74. Zuckerberg, *supra* note 5.
75. Kaye, *supra* note 11.
76. *Id*.
77. As noted above, Wikimedia, Reddit, and Twitch are welcome exceptions to this. See *supra* note 12.
78. Louise Matsakis, *Twitter Trust and Safety Advisers Say They're Being Ignored*, Wired, Aug. 23, 2019, https://perma.cc/B3C3-WJZ8.
79. Facebook, *What is the Facebook Safety Advisory Board and What Does this Board Do?*, https://perma.cc/G5M3-ASCK.

However in May 2020, Facebook took a significant new step in forming an Oversight Board to "review Facebook's most challenging content decisions—focusing on important and disputed cases."[80] Notably, the board will have power to override Facebook's moderation decisions, thus shouldering responsibility for decisions in those difficult cases. The board will not have authority to review the rules themselves, only individual decisions in which the rules were applied to remove content; nor will it have capacity to review any more than a tiny proportion of Facebook's millions of weekly takedown decisions.[81] The Oversight Board Charter does allow the board to issue policy recommendations – independently or at Facebook's request – and obligates Facebook to respond publicly within 30 days.[82] Board members might also choose to opine on the rules, in their written explanations of board decisions.

The ARTICLE 19 organization has proposed the establishment of an international "Social Media Council" (or national councils) with a significantly broader ambit than that of Facebook's Oversight Board: "The Council could elaborate ethical standards specific to the online distribution of content and cover topics such as terms and conditions, community guidelines and the content regulation practices of social media companies."[83] That council might advise multiple OSPs, as ARTICLE 19 envisions it.[84]

And in Germany, OSPs are now invited by law to consult outsiders regarding content moderation, where the purpose is to comply with German law. The Network Enforcement Act of 2017[85] is known as NetzDG, from its shortened German title, Netzwerkdurchsetzungsgesetz, but is often referred to as Germany's hate speech law though it doesn't in fact prohibit hate speech. The law requires OSPs to remove content within 24 hours if it is "manifestly unlawful" under any of 22 provisions of the German penal code and also allows them to recruit independent advisors, usually lawyers, to help them "self-regulate": to make decisions that comply with the law in difficult cases. "Under these NetzDG partnerships, committees consisting of three lawyers will provide a legal opinion on the content they receive within seven days. Tech companies will continue to do most takedowns by themselves. The partnership committees would only receive about 5–10 'high-profile' cases per month."[86]

In my view, international advisors cannot adequately contend with hate speech and other harmful forms of content, since they necessarily lack knowledge of the relevant social and political context. Also, they cannot be representative of the relevant users. OSPs should therefore recruit users to contribute both to rulemaking and to rule enforcement, on national or even local levels. This is essential for properly handling hate speech, since so much of that content can be properly understood only by those who know the detailed social, linguistic, and political context in which hate

80. Facebook, *Oversight Board Charter September 2019*, https://perma.cc/2Z3Q-K2EH.
81. Ibid. *See also* ARTICLE 19, *Facebook Oversight Board; Recommendations for Human Rights-Focused Oversight*, Mar. 29, 2019, https://perma.cc/3QSF-U2C4.
82. Facebook, *supra* note 80, at 8.
83. ARTICLE 19, Self-Regulation and "Hate Speech" on Social Media Platforms (2018), 20, *available at* https://www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-%E2%80%98hate-speech%E2%80%99-on-social-media-platforms_March2018.pdf
84. Ibid. at 21
85. Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act), Bundesministerium der Justiz und für Verbraucherschutz, July 12, 2017, https://perma.cc/R4DM-MBZ4.
86. Echikson and Knodt, *supra* note 8.

speech is made or spread. Local advisors (like their national or international counterparts) would need training in how platforms function technically, and in how to adjudicate. It would also be important, in forming local or national advisory bodies, to avoid "capture by ill-intentioned governments or groups," as Kaye points out.

National or local bodies would be able to guide platforms in adapting enforcement of their rules to their cultural and political contexts, and in tweaking the rules to conform to local social norms (as long as those do not violate international human rights law). Most OSPs insist that they maintain a single, uniform set of rules for the world (or for all countries in which they operate)—which ostensibly means they enforce the same rule against the depiction of nudity in Sweden and in Saudi Arabia. This further distances users and their own norms from the companies' rules, which should instead be adaptable to some extent.

PROPOSAL 5:

OSPs should test an array of techniques for enforcing platform rules, not only deleting content and deleting accounts.

Most efforts to diminish hateful content online use only one technique: removing it or removing the accounts from which it was posted. Takedown, as it is known in the industry, is essential for some types of egregious and/or illegal content such as child sexual exploitation, but in general it is only a stopgap, and a losing game at that, since new content is posted at a staggering rate. Moreover, removing content after it is posted is reactive, not preventive. It is roughly like pursuing food safety by removing harmful food from the market, without preventing new cases of adulteration or poisoning.[87]

The problem of "hate speech" online should be seen not simply as a matter of enforcing law or rules, but as a challenge to public welfare that requires behavior change, namely building norms of tolerance and civility.

Removing content before it is posted is a tempting alternative. Some platforms, like YouTube, already automatically detect violative content and remove it immediately after its posted; from October to December 2019, YouTube removed roughly 3.4 million videos before a single user had watched them.[88] This poses a problem for two reasons: First, hate speech cannot be automatically detected without a large margin of error.[89] Second, such removals, tantamount to prior censorship, are such strong and speech-repressive measures that they should be undertaken only with the greatest of caution, if at all, and with robust oversight from experts outside the companies that might practice it.

87. J. Nathan Matias, *A Toxic Web: What the Victorians Can Teach Us About Online Abuse,* The Guardian, Apr, 18, 2016, https://perma.cc/9V28-LE4T.
88. Google, *YouTube Community Guidelines Enforcement,* Google Transparency Report, https://perma.cc/4XCK-X2X2.
89. Saleem et al., *supra* note 38; *see also* Alba, *supra* note 30.

Already, OSPs are removing millions of posts in order to enforce their own rules, but users continue to post harmful content faster than the companies take it down. To catch up without resorting to massive automatic removal of hate speech, which could impinge severely on freedom of expression because hate speech is so difficult to detect reliably, companies need methods to persuade people not to post harmful content in the first place.

Some internet companies and researchers have begun to test and study alternate methods. These rely on an important but overlooked insight: that not all those who produce hateful content are extremists or incorrigible "trolls." Some are occasional offenders who behave better offline, and may be susceptible to online interventions. Alternative methods of enforcement for those who are not chronic producers of "hate speech" can include preventing users from posting for specific periods of time after they break a rule (some companies, such as Twitter, already do this), or requiring them to take a short online course on the rules against hate speech.

Widespread alarm about vicious content online should be channeled into new opportunities to define and reinforce norms of discourse and to learn, by means of rigorous research, how to influence behavior. This is not unrealistic: public concern has helped drive major behavior change to protect people from harm, such as wearing seat belts in motor vehicles or the decline in smoking. Even though some people continue to transgress such norms, the majority has become compliant, keeping themselves and others safer. Norms for online discourse can be greatly improved, even without eliminating "hate speech," if such norms are embraced by a critical mass of people.

Chronic offenders should be tackled differently, of course: with criminal law and prosecution where relevant, and with muscular enforcement of platform rules. Here, too, there are options that have not been sufficiently explored, such as preventing offenders from monetizing "hate speech" and other harmful or offensive content (as YouTube does[90]), setting limits on both organic and paid sharing of content (WhatsApp has tried limiting the number of groups to which one user can share a piece of content[91]), or withdrawing users' control of online spaces. For example, some Facebook pages and YouTube channels have become highly influential, with hundreds of thousands of followers, and the users who control them as administrators can delete any comments they don't like. Where user/administrators use this privilege to promote appalling views to a large number of followers[92] and to suppress dissent by anyone else, the platforms could rescind their power to do so.

On some platforms, users also have significant tools for controlling their own experience and keeping out content they don't want to see, sometimes thanks to applications built by third-party developers, and sometimes using features that the companies provide. Facebook, for example, allows users to hide posts that contain certain keywords[93], and Twitter's options to "mute" and "block" accounts can be augmented by third-party tools that allow users to share their lists of blocked accounts with

90. Google, *Advertiser-Friendly Content Guidelines*, YouTube Help, https://perma.cc/C5XX-NJK9
91. WhatsApp Inc., *More changes to forwarding,* Whatsapp Blog, Jan. 21, 2019, https://perma.cc/9VAW-BM5A.
92. For example, pages run by extreme anti-Muslim monks in Myanmar *e.g. see* Christina Fink, *Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar*, Columbia Journal of International Affairs, September 17 2018, https://perma.cc/H7HV-S6ZN.
93. Shruthi Muraleedharan, *Keyword Snooze: A New Way to Help Control Your News Feed*, Facebook Newsroom, June 27, 2019, https://perma.cc/PT4P-MSBH.

others.[94]

The best responses for countering harmful speech online will be tailored, as much as possible, to types of content, to the audiences they reach, and to the social, cultural, and historical circumstances in which they circulate. When platforms try new methods, they should rigorously test their effects, and publish the results.

PROPOSAL 6:
OSPs should communicate their rules to users more clearly and more effectively.

For almost everyone outside the companies that make and apply them, platform rules are arcane and obscure. This forecloses even the possibility of basic features of democratic governance: that people take part in debating the rules, revising them, adapting them to fit their own normative or cultural contexts, defining categorical boundaries of prohibited content, and explaining the rules to others.[95] Many of these practices would be difficult to implement on massive social platforms as they are now constructed, but that's no reason to preclude them. Internet platforms will evolve and be replaced by other models, and even on existing platforms, some scholars are testing intriguing methods to allow users to participate in governance, such as Jenny Fan and Amy Zhang's "digital juries."[96]

In the meantime, it is a dangerous precedent for most of the world to become habituated to largely invisible systems of private censorship. Moreover, making the systems more visible is not as difficult as it may seem. It would yield a variety of benefits and can be accomplished without producing collateral harm. As Tarleton Gillespie argues, "articulating the rules is the clearest opportunity for the platforms to justify their moderation efforts as legitimate."[97]

Disclosure of the rules, together with explanations of how they are applied to user requests for takedown, can also provide a sense of procedural justice that is now sorely lacking. Users of social media platforms often complain that when they report objectionable content, the response they receive from platforms says only that their request has been denied or accepted, without reference to any particular rule.[98]

---

94. Block Together: A web app intended to help cope with harassment and abuse on Twitter, *see* Block Together, https://blocktogether.org, archived at https://perma.cc/2325-DWM3.
95. OSPs debate and revise rules internally, of course, and sometimes use the language of democratic process to describe their efforts, such as the twice-monthly "mini legislative sessions" of Facebook staff, described by Monika Bickert, VP of Consumer Operations. Conference notes on file with the author; *see also* Alexis Madrigal, Inside Facebook's Fast-Growing Content-Moderation Effort, The Atlantic, Feb. 7, 2018, https://perma.cc/UT3G-EP8D.
96. Jenny Fan and Amy X. Zhang, *Digital Juries: A Civics-Oriented Approach to Platform Governance*, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (April 2020), https://perma.cc/3LS2-N2BQ
97. Gillespie, *supra* note 4, 45.
98. In response to such complaints, some OSPs, including YouTube and Twitter, have begun to explain their decisions whether or not to remove content in response to individual requests for takedown.

There is also considerable evidence that people who are familiar with rules are more likely to follow them.[99] Since OSP rules are designed to prevent or at least discourage a variety of serious individual and collective harms, it would be of major social benefit if fewer internet users broke the rules and/or did so less often.

OSPs can easily make more users aware of their outward-facing content regulations. They typically present those rules in thousands of words of fine print, buried in their terms of service, which the vast majority of users never read. Many do not even know they exist.[100]

In a 2017 study, every one of 543 college students in a laboratory experiment clicked the "Join" button for a new social network, unwittingly consenting in paragraph 2.3.1 of the terms of service to give the network not only their data but also their future first-born child.[101] In an observational study of online behavior, fewer than 0.2% of online software buyers spent even one second looking at the terms of service before accepting them.[102] For users of OSPs, accepting these terms is, effectively, a contract of adhesion in which content moderation rules are buried.

Writing the rules in clear, simple language and obliging users to read them is a mild and uncomplicated intervention that is very unlikely to do any harm, and can make people more likely to follow the rules. Prof. J. Nathan Matias worked with moderators on the large subreddit r/science to test for this effect. When the rules of the subreddit were pinned to the top of each comment thread, those who commented were significantly less likely to break the rules. [103]

Other efforts to improve online norms of behavior by making rules visible give some early basis for cautious optimism. There have been reports of successful online behavior modification: by Facebook, to teach users to resolve grievances successfully with one another;[104] by the online gaming company Riot Games, to decrease "toxic" comments by players of League of Legends, a game played by millions around the world;[105] and even as far back as the 1990s, by the Massachusetts Institute of Technology's then-director of academic computing, to reduce online harassment of students.[106] These experiences should provide a trove of information, but thus far the findings have not been published in sufficient detail to permit replication or statistical analysis. It is essential to build up an accessible and rigorous body of knowledge about ways to diminish harmful online behavior, e.g. by

---

99. *See* e.g. M. E. Tankard & E. L. Paluck, *supra* note 20.
100. Anecdotal surveys by the author, in which US college students blinked in confusion when asked if they had ever read the community guidelines or content rules of any platform they used. Many students said they were not aware that such rules existed, or were available to read.
101. David Berreby, *Click to Agree with What? No One Reads Terms of Service, Studies Confirm,* The Guardian, Mar. 3, 2017, https://perma.cc/FVB3-DMEA.
102. Yannis Bakos, Florencia Marotta-Wurgler, David R. Trossen, *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts,* 43.1 J. Legal Stud. 1 (2014)
103. J. Nathan Matias, Governing Human and Machine Behavior in an Experimenting Society (2017) (unpublished PhD thesis, Massachusetts Institute of Technology), https://perma.cc/B5RT-9V7W.
104. Jason Marsh, *Can Science Make Facebook More Compassionate*, Greater Good Magazine, July 25, 2012, https://perma.cc/68XU-2T82.
105. Brendan Maher, *Can a Video Game Company Tame Toxic Behavior?*, Nature, Mar. 30, 2016, https://perma.cc/PXU6-28PE.
106. Gregory A. Jackson, *Promoting Network Civility at MIT: Crime & Punishment, or the Golden Rule?,* 75.3 Educational Record 29 (1994), https://perma.cc/65E7-765V.

communicating the rules clearly.

When they began the discussion site Parlio in 2014 with the goal of fostering civil public conversation among people who strongly disagree with one another, Wael Ghonim and his co-founders required new users to read and accept a simple set of rules, presented one at a time in a relatively large font and few words, so it was almost impossible to ignore them. So many users remained civil that the platform staff found itself with very few moderation dilemmas to discuss at their weekly meetings (They did note that their users may have been a disproportionately civil sample of the population even before they joined Parlio.[107]) Parlio also posted a brief statement emphasizing the new site's focus on—and enforced demand for—civility.[108]

Of course, not all users will be swayed by such interventions. Many will continue to ignore rules, some will be unable to understand them, and highly motivated trolls and other producers of harmful content may even be inspired to work harder to flout them. Some of those producers are not only highly motivated but vigorously supported and/or employed by governments in many countries, such as Russia, China, and Brazil.[109] However, there is evidence that, at least on some platforms, a majority of the hateful content is produced not by chronic trolls or bad actors but by users who do so only occasionally.

In internal research at the company Riot Games, Jeffrey Lin found that only about 1% of League of Legends players were consistently producing what he called toxic content, and that they were responsible for less than 5% of such content on the platform.[110] The rest was produced by intermittent violators who were usually civil. If a critical mass of those users becomes familiar with the rules, there may be a net favorable effect as the rules become better accepted as robust norms of behavior.

It is an old and familiar process, after all: many of the major improvements in human life of the last decades are the result of behavior change driven by shifts in social norms, such as not smoking, wearing seatbelts, boiling unsanitary water before giving it to infants, and so on. Though some people fail to comply and some vigorously continue to violate norms, most enjoy both the individual and collective benefits. Further, in this case making the rules transparent may plant seeds for other forms of effective engagement by users.

---

107. Interview on file with author, 2015.

108. The text of the statement: "Be curious, open-minded, and civil. We want you to share opinions and experiences that strengthen the community's collective intelligence. We believe diversity of thought is a virtue, and we're here to learn new perspectives; not to win arguments. We are trying to define a new type of network. One void of Internet-trolling, where we can create a community of trust and respect that expands our horizons. Parlio values dissent, but above all else, civility."

109. Samuel C. Woolley and Philip N. Howard, Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media (2018). https://doi.org/10.1093/oso/9780190931407.001.0001

110. Maher, *supra* note 105.

PROPOSAL 7:
OSPs should allow external oversight of enforcement mechanisms.

OSPs have been rapidly expanding, speeding up, and automating their content moderation systems, and will continue to do so under increasing pressure from users and especially from governments. Yet no one outside the companies has anything more than a vague and anecdotal sense of which content is being taken down and which content remains online. The photograph of the Vietnamese girl running while napalm burned her body, for example, galvanized extensive public discussion about Facebook's rules and especially about how they are enforced, but it is just one piece of content among approximately one million such pieces that Facebook removes every day, not including content that it classifies as spam.[111] Though it is important to publicize rules, it is also critical to understand how they are being put into practice or enforced. In order to protect freedom of expression, it is therefore essential to construct a mechanism for oversight.

Such a mechanism could have serious implications for user privacy, since it would be difficult to review actual moderation decisions without seeing examples of real content, posted from real accounts. Among other concerns, tech companies also worry that releasing such data might expose them to prosecution for failing to remove content that governments deem illegal.

User data could be protected in one of two ways. First, data could be released only under a rigorous system of the type commonly used in the social sciences for sensitive data that includes private information. Such data can be accessed only under strictly controlled conditions, and only by researchers who have been vetted in advance. In this case, outside reviewers would be forbidden to release any actual content, or information about individual users or accounts. Second, data could be released exclusively to independent boards that would be set up for that purpose and vetted in advance.

Finally, some testing can be done even without data provided by OSPs, though researchers might then violate an OSP's terms of service which often prohibit "scraping" of data. Some countries also have laws against this, like the U.S. Computer Fraud and Abuse Act which criminalizes "unauthorized access" to a website.[112] In any case, review of moderation systems should become standard procedure, just as food safety inspectors visit restaurant kitchens on an intermittent but regular basis.

111. Facebook, Community Standards Enforcement Report, Facebook Transparency Report (May 2020), https://perma.cc/7NZK-6SPD.

112. See e.g. Brian Z. Mund, *Comment, Protecting Deceptive Academic Research Under the Computer Fraud and Abuse Act*, 37 Yale Law & Policy Review 385 (2018), https://perma.cc/DPG6-2PRJ.

APPENDIX:
OSP HATE SPEECH POLICIES

The following are excerpts from various internet companies' policies on hate speech or hateful conduct, or relevant portions of their Terms of Service or other similar documents. In some cases, sections not related to hateful speech have been omitted for clarity and brevity.

Facebook[113]

We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion, and in some cases, may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.

Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others. In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. People sometimes express contempt in the context of a romantic break-up. Other times, they use gender-exclusive language to control membership in a health or positive support group, such as a breastfeeding group for women only. In all of these cases, we allow the content but expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.

We allow humour and social commentary related to these topics. In addition, we believe that people are more responsible when they share this kind of commentary using their authentic identity.

Do not post:

Tier 1

Content targeting a person or group of people (including all subsets except those described as having carried out violent crimes or sexual offences) on the basis of their aforementioned protected characteristic(s) or immigration status with:

- Violent speech or support in written or visual form
- Dehumanizing speech such as reference or comparison to:
    - Insects
    - Animals that are culturally perceived as intellectually or physically inferior
    - Filth, bacteria, disease and feces
    - Sexual predator

113. Facebook, *Community Standards,* sec. 12. "Hate Speech," https://perma.cc/LML9-9UX8.

- Subhumanity
- Violent and sexual criminals
- Other criminals (including but not limited to "thieves," "bank robbers," or saying "all [protected characteristic or quasi-protected characteristic] are "criminals'")
- Statements denying existence

• Mocking the concept, events or victims of hate crimes, even if no real person is depicted in an image

• Designated dehumanizing comparisons, generalizations, or behavioral statements (in written or visual form)- that include:

- Black people and apes or ape-like creatures
- Black people and farm equipment
- Jewish people and rats
- Muslim people and pigs
- Muslim person and sexual relations with goats or pigs
- Mexican people and worm like creatures
- Women as household objects or referring to women as property or 'objects'
- Transgender or non-binary people referred to as 'it'

Tier 2

Content targeting a person or group of people on the basis of their protected characteristic(s) with:

• Generalisations that state inferiority (in written or visual form) in the following ways:

- Physical deficiencies are defined as those about:
  - Hygiene, including but not limited to: filthy, dirty, smelly
  - Physical appearance, including but not limited to: ugly, hideous
- Mental deficiencies are defined as those about:
  - Intellectual capacity, including but not limited to: dumb, stupid, idiots
  - Education, including but not limited to: illiterate, uneducated
  - Mental health, including but not limited to: mentally ill, retarded, crazy, insane
- Moral deficiencies are defined as those about:
  - Culturally perceived negative character trait, including but not limited to: coward, liar, arrogant, ignorant
  - Derogatory terms related to sexual activity, including but not limited to: whore, slut, perverts
- Other statements of inferiority, which we define as:
  - Expressions about being less than adequate, including but not limited to: worthless, useless

    - Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females."
    - Expressions about deviating from the norm, including but not limited to: freaks, abnormal
- Expressions of contempt or their visual equivalent, which we define as:
    - Self-admission to intolerance on the basis of a protected characteristics, including but not limited to: homophobic, islamophobic, racist
    - Expressions that a protected characteristic shouldn't exist
    - Expressions of hate, including but not limited to: despise, hate
    - Expressions of dismissal, including but not limited to: don´t respect, don't like, don´t care for
- Expressions of disgust or their visual equivalent, which we define as:
    - Expressions that suggest the target causes sickness, including but not limited to: vomit, throw up
    - Expressions of repulsion or distaste, including but not limited to: vile, disgusting, yuck
- Cursing, defined as:
    - Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole
    - Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker
    - Terms or phrases calling for engagement in sexual activity, or contact with genitalia or anus, or with feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit

Tier 3

Content targeting a person or group of people on the basis of their protected characteristic(s) with any of the following:

- Calls for segregation
- Explicit Exclusion which includes but is not limited to "expel" or "not allowed".
- Political Exclusion defined as denial of right to political participation.
- Economic Exclusion defined as denial of access to economic entitlements and limiting participation in the labour market,
- Social Exclusion defined as including but not limited to denial of opportunity to gain access to spaces (incl. online) and social services.

We do allow criticism of immigration policies and arguments for restricting those policies.

Content that describes or negatively targets people with slurs, where slurs are defined as words

commonly used as insulting labels for the above-listed characteristics.

Twitter[114]

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

Rationale
Twitter's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right—we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives.

We recognize that if people experience abuse on Twitter, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. This includes; women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities. For those who identity with multiple underrepresented groups, abuse may be more common, more severe in nature and have a higher impact on those targeted.

We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals with abuse based on protected category.

If you see something on Twitter that you believe violates our hateful conduct policy, please report it to us.

When this applies
We will review and take action against reports of accounts targeting an individual or group of people with any of the following behavior, whether within Tweets or Direct Messages.

Violent threats
We prohibit content that makes violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm,

114. Twitter Inc., *Hateful Conduct Policy,* Twitter Help Center, https://perma.cc/8MCM-FAAS.

where an individual could die or be significantly injured, e.g., "I will kill you."

Note: we have a zero tolerance policy against violent threats. Those deemed to be sharing violent threats will face immediate and permanent suspension of their account.

Wishing, hoping or calling for serious harm on a person or group of people
We prohibit content that wishes, hopes, promotes, or expresses a desire for death, serious and lasting bodily harm, or serious disease against an entire protected category and/or individuals who may be members of that category. This includes, but is not limited to:

- Hoping that someone dies as a result of a serious disease, e.g., "I hope you get cancer and die."
- Wishing for someone to fall victim to a serious accident, e.g., "I wish that you would get run over by a car next time you run your mouth."
- Saying that a group of individuals deserve serious physical injury, e.g., "If this group of protesters don't shut up, they deserve to be shot."

References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims
We prohibit targeting individuals with content that references forms of violence or violent events where a protected category was the primary target or victims, where the intent is to harass. This includes, but is not limited to sending someone:

- media that depicts victims of the Holocaust;
- media that depicts lynchings.

Inciting fear about a protected category
We prohibit targeting individuals with content intended to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., "all [religious group] are terrorists."

Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone
We prohibit targeting individuals with repeated slurs, tropes or other content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes about a protected category. This includes targeted misgendering or deadnaming of transgender individuals.

We also prohibit the dehumanization of a group of people based on their religion, age, disability, or serious disease.

Hateful imagery
We consider hateful imagery to be logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, gender identity or ethnicity/national origin. Some examples of hateful imagery include, but are not limited to:

- symbols historically associated with hate groups, e.g., the Nazi swastika;
- images depicting others as less than human, or altered to include hateful symbols, e.g.,

altering images of individuals to include animalistic features; or

- images altered to include hateful symbols or references to a mass murder that targeted a protected category, e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust.

Media depicting hateful imagery is not permitted within live video, account bio, profile or header images. All other instances must be marked as sensitive media. Additionally, sending an individual unsolicited hateful imagery is a violation of our abusive behavior policy.

Do I need to be the target of this content for it to be a violation of the Twitter Rules?
Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.

When we review this type of content, it may not be clear whether the intention is to abuse an individual on the basis of their protected status, or if it is part of a consensual conversation. To help our teams understand the context, we sometimes need to hear directly from the person being targeted to ensure that we have the information needed prior to taking any enforcement action.

Note: individuals do not need to be a member of a specific protected category for us to take action. We will never ask people to prove or disprove membership in any protected category and we will not investigate this information.

Consequences
Under this policy, we take action against behavior that targets individuals or an entire protected category with hateful conduct, as described above. Targeting can happen in a number of ways, for example, mentions, including a photo of an individual, referring to someone by their full name, etc.

Under this policy, we take action against behavior that targets individuals or an entire protected category with hateful conduct, as described above. Targeting can happen in a number of ways, for example, mentions, including a photo of an individual, referring to someone by their full name, etc.

When determining the penalty for violating this policy, we consider a number of factors including, but not limited to the severity of the violation and an individual's previous record of rule violations. For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent account suspension. If an account is engaging primarily in abusive behavior, or is deemed to have shared a violent threat, we will permanently suspend the account upon initial review.

YouTube[115]

Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes:

- **Age**
- **Caste**
- **Disability**
- **Ethnicity**
- **Gender Identity and Expression**
- **Nationality**
- **Race**
- **Immigration Status**
- **Religion**
- **Sex/Gender**
- **Sexual Orientation**
- **Victims of a major violent event and their kin**
- **Veteran Status**

If you see content that violates this policy, please report it. If you have found multiple videos, comments, or a user's entire channel that you wish to report, please visit our reporting tool, where you will be able to submit a more detailed complaint.

What this means for you
If you're posting content
Don't post content on YouTube if the purpose of that content is to do one or more of the following.

- Encourage violence against individuals or groups based on any of on the attributes noted above. We don't allow threats on YouTube, and we treat implied calls for violence as real threats. You can learn more about our policies on threats and harassment.
- Incite hatred against individuals or groups based on any of the attributes noted above.

Other types of content that violates this policy
- Dehumanizing individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease, or any other non-human entity.
- Praise or glorify violence against individuals or groups based on the attributes noted above.
- Use of racial, religious or other slurs and stereotypes that incite or promote hatred based on any of the attributes noted above. This can take the form of speech, text, or imagery promoting these

115. Google, *Hate Speech Policy*, YouTube Help, https://perma.cc/6826-ZP3J.

stereotypes or treating them as factual.

- Claim that individuals or groups are physically or mentally inferior, deficient, or diseased based on any of the attributes noted above. This includes statements that one group is less than another, calling them less intelligent, less capable, or damaged.
- Allege the superiority of a group over those with any of the attributes noted above to justify violence, discrimination, segregation, or exclusion.
- Conspiracy theories ascribing evil, corrupt, or malicious intent to individuals or groups based on any of the attributes noted above.
- Call for the subjugation or domination over individuals or groups based on any of the attributes noted above.
- Deny that a well-documented, violent event took place.
- Attacks on a person's emotional, romantic and/or sexual attraction to another person.
- Content containing hateful supremacist propaganda including the recruitment of new members or requests for financial support for their ideology.
- Music videos promoting hateful supremacism in the lyrics, metadata, or imagery.

Educational content
We may allow content that includes hate speech if the primary purpose is educational, documentary, scientific, or artistic in nature. This is not a free pass to promote hate speech. Examples include:

- A documentary about a hate group: Educational content that isn't supporting the group or promoting ideas would be allowed. A documentary promoting violence or hatred wouldn't be allowed.
- A documentary about the scientific study of humans: A documentary about how theories have changed over time, even if it includes theories about the inferiority or superiority of specific groups, would be allowed because it's educational. We won't allow a documentary claiming there is scientific evidence today that an individual or group is inferior or subhuman.
- Historical footage of an event, like WWII, which doesn't promote violence or hatred.

This policy applies to videos, video descriptions, comments, live streams, and any other YouTube product or feature. For educational content that includes hate speech, this context must appear in the images or audio of the video itself. Providing it in the title or description is insufficient.

Examples
Here are examples of hate speech not allowed on YouTube.

- "I'm glad this [violent event] happened. They got what they deserved [referring to persons with the attributes noted above]."
- "[Person with attributes noted above] are dogs" or "[person with attributes noted above] are like animals."

More examples

- "Get out there and punch a [person with attributes noted above]"
- "Everyone in [groups with attributes noted above] are all criminals and thugs."
- "[Person with attributes noted above] is scum of the earth."
- "[People with attributes noted above] are a disease."
- "[People with attributes noted above] are less intelligent than us because their brains are smaller."
- "[Group with any of the attributes noted above] threaten our existence, so we should drive them out at every chance we get."
- "[Group with any of the attributes noted above] has an agenda to run the world and get rid of us."
- "[Attribute noted above] is just a form of mental illness that needs to be cured."
- "[Person with any of the attributes noted above] shouldn't be educated in schools because they shouldn't be educated at all."
- "All of the so-called victims of this violent event are actors. No one was hurt, and this is just a false flag."
- "All of the 'so-called victims' of this are actors. No one was hurt."
- Shouting "[people with attributes noted above] are pests!" at someone regardless of whether the person does or does not have the alleged attributes.
- Video game content which has been developed or modified ("modded") to promote violence or hatred against a group with any of the attributes noted above.

Please remember these are just some examples, and don't post content if you think it might violate this policy.

What happens when content violates this policy

If your content violates this policy, we'll remove the content and send you an email to let you know. If this is your first time violating our Community Guidelines, you'll get a warning with no penalty to your channel. If it's not, we'll issue a strike against your channel. If you get 3 strikes, your channel will be terminated.

If we think your content comes close to hate speech, we may limit YouTube features available for that content.

Instagram[116]

Instagram is a reflection of our diverse community of cultures, ages, and beliefs. We've spent a lot of time thinking about the different points of view that create a safe and open environment for everyone.

We created the Community Guidelines so you can help us foster and protect this amazing community. By using Instagram, you agree to these guidelines and our Terms of Use. We're committed to these guidelines and we hope you are too. Overstepping these boundaries may result in deleted content, disabled accounts, or other restrictions.

Follow the law.
Instagram is not a place to support or praise terrorism, organized crime, or hate groups.

Respect other members of the Instagram community.
We want to foster a positive, diverse community. We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages. We do generally allow stronger conversation around people who are featured in the news or have a large public audience due to their profession or chosen activities.

It's never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your intent clearly.

Serious threats of harm to public and personal safety aren't allowed. This includes specific threats of physical harm as well as threats of theft, vandalism, and other financial harm. We carefully review reports of threats and consider many things when determining whether a threat is credible.

Be thoughtful when posting newsworthy events.
We understand that many people use Instagram to share important and newsworthy events. Some of these issues can involve graphic images. Because so many different people and age groups use Instagram, we may remove videos of intense, graphic violence to make sure Instagram stays appropriate for everyone.

We understand that people often share this kind of content to condemn, raise awareness or educate. If you do share content for these reasons, we encourage you to caption your photo with a warning about graphic violence. Sharing graphic images for sadistic pleasure or to glorify violence is never allowed.

116. Instagram, Inc., *Community Guidelines,* Instagram Help Center, https://perma.cc/W7NM-D28H.

Tumblr[117]

What Tumblr is for:

Tumblr celebrates creativity. We want you to express yourself freely and use Tumblr to reflect who you are, and what you love, think, and stand for.

What Tumblr is not for:

• **Terrorism.** We don't tolerate content that promotes, encourages, or incites acts of terrorism. That includes content which supports or celebrates terrorist organizations, their leaders, or associated violent activities.

• **Hate Speech.** Don't encourage violence or hatred. Don't post content for the purpose of promoting or inciting the hatred of, or dehumanizing, individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, veteran status, sexual orientation, disability or disease. If you encounter content that violates our hate speech policies, please report it.

Keep in mind that a post might be mean, tasteless, or offensive without necessarily encouraging violence or hatred. In cases like that, you can always block the person who made the post—or, if you're up for it, you can express your concerns to them directly, or use Tumblr to speak up, challenge ideas, raise awareness or generate discussion and debate.

• **Violent Content and Threats, Gore and Mutilation.** Don't post content which includes violent threats toward individuals or groups - this includes threats of theft, property damage, or financial harm. Don't post violent content or gore just to be shocking. Don't showcase the mutilation or torture of human beings, animals (including bestiality), or their remains. Don't post content that encourages or incites violence, or glorifies acts of violence or the perpetrators.

• **Harassment.** Don't engage in targeted abuse, bullying, or harassment. Don't engage in the unwanted sexualization or sexual harassment of others. If someone is sending you unwanted messages, or reblogging your posts in an abusive way, we encourage you to be proactive. Report them, and block the hell out of them. And if someone blocks you, don't attempt to circumvent the block feature or otherwise try to communicate with them.

If we conclude that you are violating these guidelines, you may receive a notice via email. If you don't explain or correct your behavior, we may take action against your account. We do our best to ensure fair outcomes, but in all cases we reserve the right to suspend accounts, or remove content, without notice, for any reason, but particularly to protect our services, infrastructure, users, and community. We reserve the right to enforce, or not enforce, these guidelines in our sole discretion, and these guidelines don't create a duty or contractual obligation for us to act in any particular manner.

117. Tumblr, *Community Guidelines*, Jan. 23, 2020, https://perma.cc/C26L-PAQF.

Microsoft

Microsoft Services Agreement[118]

3. Code of Conduct.
    a. By agreeing to these Terms, you're agreeing that, when using the Services, you will follow these rules:
        vii. Don't engage in activity that is harmful to you, the Services, or others (e.g., transmitting viruses, stalking, posting terrorist content, communicating hate speech, or advocating violence against others).

b. Enforcement. If you violate these Terms, we may stop providing Services to you or we may close your Microsoft account. We may also block delivery of a communication (like email, file sharing or instant message) to or from the Services in an effort to enforce these Terms or we may remove or refuse to publish Your Content for any reason. When investigating alleged violations of these Terms, Microsoft reserves the right to review Your Content in order to resolve the issue. However, we cannot monitor the entire Services and make no attempt to do so.

Community Standards for Xbox[119]

We built Xbox Live for people like you—for players from all walks of life, everywhere in the world, who all want the same thing: a place to play and have fun. We need your help keeping the Xbox online community safe and fun for everyone.

While the Code of Conduct section of the Microsoft Services Agreement applies to all Microsoft products, Xbox Live offers so many ways to interact with others that it benefits from an additional level of explanation.

To this end, we've created the following community standards for Xbox. Consider these standards a roadmap for contributing to this incredible, globe-spanning community. Remember: Xbox Live is your community. We all bring something unique, and that uniqueness is worth protecting.

Whether you're brand new to gaming or have been playing for decades, we need you to be stewards of this place, to protect each other even as you compete. Because when everyone plays, we all win.

Our Shared Values

The spirit of Xbox lives in our values, which are key to sustaining a vibrant and welcoming community. Living these values every time we play shows the world the unifying power of gaming.

- Gaming can be enjoyed by all
- Creativity powers community
- Competition is best when it's fair
- Helping others makes all of us stronger
- Hate has no place here

Conduct

118. Microsoft, *Microsoft Services Agreement*, July 1, 2019, https://perma.cc/VJG2-F9D5.
119. Microsoft, *Community Standards for Xbox*, https://perma.cc/X95L-F2LN

Some parts of the internet don't have rules—and the Xbox online community isn't one of them. Yes, Xbox Live is, in a meaningful sense, your gaming network. But it belongs to millions of others, too. You deserve a place to be yourself with confidence, free from bullying, hatred, and harassment—and so does every other player. So it's important to treat others as they would like to be treated.

Remember:
• Win or lose, be a good sport
• Did someone have a great game? Let them know!
• You are the community
• A little bit of trash talk is okay, but keep it clean
• No one likes trolling, so don't do it

Content
The gamertags, gamerpics, screenshots, game clips, and other posts you make on Xbox can be a great way to show off what's meaningful to you. We encourage all players to be themselves and show off what they like, what makes them laugh, or what makes them amazing. But this sharing can't come at the expense of other players' positive experiences.

Remember:
• Use your skills and creativity to add informative, helpful, funny, or interesting content that contributes positively to our vibrant and diverse community
• Content you post on Xbox needs to suit a wide audience
• Context is important, and mature content that makes sense in a game might not be appropriate elsewhere on Xbox
• Not everyone has the same likes or dislikes as you, so think twice about saying something hurtful about someone else's content, playing style, or choices

Standards
If you've seen the Microsoft Services Agreement, the following rules probably look familiar. They may sound a bit like legalese, but bear with us—upholding these standards is critical to maintaining a community where everyone can have fun! People differ about what seems fun, and conflicts sometimes occur. But while plenty of conflicts can be worked out between players, there are nevertheless some things we just can't tolerate.

In each section you'll find examples showing how the Microsoft Services Agreement's Code of Conduct relates to Xbox Live.

ii. Do your part to keep everyone safe
To keep Xbox Live a place where everyone can have fun, we can't allow behavior or content designed to exploit, harm, or threaten anyone – children, adults, or otherwise. When threatening, abusive, or insulting language is used against another member of our community, or the community at large, it undermines every player's ability to enjoy themselves.

For example, don't:
• Threaten someone with physical assault after an intense game
• Message other players with homophobic slurs
• Make a club grounded in ethnic hatred
• Create a Looking for Group that negatively calls out another player
• Post insults in another player's activity feed
• Respond to someone's smack talk with sexual slurs

iv. Keep your content clean
People enjoy all shapes and styles of content on Xbox. Everyone's tastes are different, and that's great! However, that doesn't mean that absolutely anything goes. To keep Xbox Live welcoming and inclusive for everyone, some content must be avoided.

Support a welcoming and inclusive community
Harassment and hate take many forms, but none have a home on Xbox. To make Xbox Live a place where everyone can hang out, and to prevent people from feeling uncomfortable or unwelcome, we all need to be stewards. This means more than just not harassing other players—it means embracing them. It means saving those unsavory jokes for people you know will enjoy them. It means taking particular care for others while you play, keeping in mind how they might interpret your content.

For example, don't:
• Make fun of other people's identities or personal traits
• Send harassing or abusive messages
• Use a club to shame other players or groups
• Start a broadcast in order to troll someone
• Flood voice chat with music during a multiplayer match
• Post game clips that will offend many others

Know the difference between trash talk and harassment
We get it—gaming can be competitive and interactions with other players can get heated. A little trash talk is an expected part of competitive multiplayer action, and that's not a bad thing. But hate has no place here, and what's not okay is when that trash talk turns into harassment.

Trash talk includes any lighthearted banter or bragging that focuses on the game at hand and encourages healthy competition. Harassment includes any negative behavior that's personalized, disruptive, or likely to make someone feel unwelcome or unsafe. To qualify as harassment, the behavior doesn't have to be drawn-out or persistent. Even a single abusive message could harm someone's experience. Know when to draw the line, when to back off. Know and respect the other player.

For example:
Acceptable trash talk includes
• Get destroyed. Can't believe you thought you were on my level.
• That was some serious potato aim. Get wrecked.
• Only reason you went positive was you spent all game camping. Try again, kid.

• Cheap win. Come at me when you can actually drive without running cars off the road.
• That sucked. Get good and then come back when your k/d's over 1.

Going too far looks like
• Get <sexual threat>. Can't believe you thought you were on my level.
• Hey <profanity>, that was some serious potato aim. Get wrecked, trash.
• Only reason you went positive was you spent all game camping. KYS, kid.
• Cheap win. Totally expected from a <racial slur>.
• You suck. Get out of my country—maybe they'll let you back in when your k/d's over 1.

Consequences
Our priority is the safety and enjoyment of everyone on Xbox Live. Content and behavior that puts players at risk or makes them feel unwelcome has no place in the Xbox online community. So, sometimes we need to step in. We're not out to punish, but rather to protect everyone's experience.

Every suspension or other corrective action aims only to show what was wrong and what can be learned from a situation. When suspensions end, we welcome players back so they can contribute to Xbox Live in positive ways. We know people make mistakes, and we believe lapses in judgment can be significant opportunities for growth.

Inappropriate conduct
If you violate Xbox community standards, you may find restrictions placed on your profile and/or device. When we suspend an Xbox profile, we restrict access to features that are most closely associated with the problematic behavior. Most commonly, this means a temporary suspension that removes one or more features for a period of time.

Temporary suspensions can include:
• Restrictions on the use of online multiplayer gaming
• Removal of the ability to send text and voice messages on Xbox
• Blocking real-time voice and text communications on Xbox
• Preventing the broadcast of live game play
• Restrictions on the use of parties and clubs

Inappropriate content
Since Xbox Live content must be appropriate for all audiences, sometimes we remove content to protect our customers. Depending on the type of content violation, this can result in our restricting certain features for the profile that created or shared the content.

Temporary suspensions can include:
• Blocks on the ability to upload game clips and screenshots to Xbox Live
• Restrictions on uploading or sharing Kinect content
• Removal of inappropriate content from Xbox Live
• Automatic assignment of a new gamertag
• Limits on the ability to share Xbox content on other social networks
• Removal of the ability to edit your Xbox profile or clubs

Repeat or severe offenses
We may permanently suspend a profile or device if we can no longer trust it due to a severe violation, or if our attempts to correct repeated negative behaviors are unsuccessful. Under permanent suspension, the owner of the suspended profile forfeits all licenses for games and other content, Gold membership time, and Microsoft account balances.

Microsoft Hate Speech Reporting Form[120]
At Microsoft, we recognize that we have an important role to play in fostering safety and civility on our hosted consumer services.

Please use this web form to report content posted or shared on Microsoft-hosted consumer services that may constitute hate speech - for example, content that advocates violence or promotes hatred based on:

- Age
- Disability
- Gender
- National or ethnic origin
- Race
- Religion
- Sexual orientation
- Gender identity

Please note that not all content that you may find offensive is considered hate speech and, in reviewing your report, Microsoft may choose to take no action.

WhatsApp[121]

Legal and Acceptable Use.
You must access and use our Services only for legal, authorized, and acceptable purposes. You will not use (or assist others in using) our Services in ways that:

- (b) are illegal, obscene, defamatory, threatening, intimidating, harassing, hateful, racially, or ethnically offensive, or instigate or encourage conduct that would be illegal, or otherwise inappropriate, including promoting violent crimes
- (c) involve publishing falsehoods, misrepresentations, or misleading statements

120. Microsoft, "Report Hate Speech Content Posted to a Microsoft Hosted Consumer Service," https://perma.cc/QBM3-MNSW.
121. WhatsApp.com, *WhatsApp Terms of Service,* Jan. 28, 2020, https://perma.cc/A5ZY-BZD9.

Pinterest[122]

Our team works hard to keep divisive, disturbing or unsafe content off Pinterest. We delete some types of content, and other stuff we just hide from public areas.

We remove hate speech and discrimination, or groups and people that advocate either. Hate speech includes serious attacks on people based on their race, ethnicity, national origin, religion, gender identity, sexual orientation, disability or medical condition. Also, please don't target people based on their age, weight, immigration or ex-military status.

We remove content used to threaten or organize violence or support violent organizations. We don't allow anything that presents a real risk of harm to people or property. We also don't want anyone making threats, organising violence or encouraging others to be violent.

Any person or group that's dedicated to causing harm to others isn't welcome on Pinterest. That includes terrorist organisations and gangs. We collaborate with industry, government and security experts to help us identify these groups.

We remove harmful advice, content that targets individuals or protected groups and content created as part of disinformation campaigns.
Don't put harmful misinformation on Pinterest.

- We don't allow advice when it has immediate and detrimental effects on a pinner's health or on public safety. This includes promotion of false cures for terminal or chronic illnesses and anti-vaccination advice.

- We don't allow misinformation about protected groups that promotes fear, hate and prejudice. This and other policies, including our hate speech guidelines, are designed to keep Pinterest a positive and welcoming environment for people of all backgrounds.

- We don't allow misinformation that attacks individuals and turns them, their families or their properties into targets of harassment or violence.

- We don't allow content that originates from disinformation campaigns targeted at Pinterest or other platforms.

- We don't allow false or misleading content that impedes the integrity of an election or an individual's or group's civic participation, including registering to vote, voting, and being counted in a census.

122. Pinterest, *Community Guidelines,* https://perma.cc/VLD9-9BG3.

Airbnb

Airbnb Community Standard: Fairness[123]

The global Airbnb community is as diverse, unique, and vibrant as the world around us. Fairness is what holds us together, what makes it possible for us to trust one another, integrate seamlessly within communities, and feel as if can we can truly belong.

Discriminatory behavior or hate speech

You should treat everyone with respect in every interaction. So, you should follow all applicable laws and not treat others differently because of their race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, disability, or serious diseases. Similarly, insulting others on these bases is not allowed.

Airbnb's Nondiscrimination Policy: Our Commitment to Inclusion and Respect[124]

Airbnb is, at its core, an open community dedicated to bringing the world closer together by fostering meaningful, shared experiences among people from all parts of the world. Our community includes millions of people from virtually every country on the globe. It is an incredibly diverse community, drawing together individuals of different cultures, values, and norms.

The Airbnb community is committed to building a world where people from every background feel welcome and respected, no matter how far they have traveled from home. This commitment rests on two foundational principles that apply both to Airbnb's hosts and guests: inclusion and respect. Our shared commitment to these principles enables every member of our community to feel welcome on the Airbnb platform no matter who they are, where they come from, how they worship, or whom they love. Airbnb recognizes that some jurisdictions permit, or require, distinctions among individuals based on factors such as national origin, gender, marital status or sexual orientation, and it does not require hosts to violate local laws or take actions that may subject them to legal liability. Airbnb will provide additional guidance and adjust this nondiscrimination policy to reflect such permissions and requirements in the jurisdictions where they exist.

While we do not believe that one company can mandate harmony among all people, we do believe that the Airbnb community can promote empathy and understanding across all cultures. We are all committed to doing everything we can to help eliminate all forms of unlawful bias, discrimination, and intolerance from our platform. We want to promote a culture within the Airbnb community—hosts, guests and people just considering whether to use our platform—that goes above and beyond mere compliance. To that end, all of us, Airbnb employees, hosts and guests alike, agree to read and act in accordance with the following policy to strengthen our community and realize our mission of ensuring

123. Airbnb, *Community Standards*, https://perma.cc/QZ4L-5DQ2
124. Airbnb, *Nondiscrimination Policy*, https://www.airbnb.com/terms/nondiscrimination_policy

that everyone can belong, and feels welcome, anywhere.

- Inclusion – We welcome guests of all backgrounds with authentic hospitality and open minds. Joining Airbnb, as a host or guest, means becoming part of a community of inclusion. Bias, prejudice, racism, and hatred have no place on our platform or in our community. While hosts are required to follow all applicable laws that prohibit discrimination based on such factors as race, religion, national origin, and others listed below, we commit to do more than comply with the minimum requirements established by law.

- Respect – We are respectful of each other in our interactions and encounters. Airbnb appreciates that local laws and cultural norms vary around the world and expects hosts and guests to abide by local laws, and to engage with each other respectfully, even when views may not reflect their beliefs or upbringings. Airbnb's members bring to our community an incredible diversity of background experiences, beliefs, and customs. By connecting people from different backgrounds, Airbnb fosters greater understanding and appreciation for the common characteristics shared by all human beings and undermines prejudice rooted in misconception, misinformation, or misunderstanding.

Specific Guidance for Hosts in the United States and European Union

Guided by these principles, our U.S. and EU host community will follow these rules when considering potential guests and hosting guests:

Race, Color, Ethnicity, National Origin, Religion, Sexual Orientation, Gender Identity, or Marital Status

Airbnb hosts may not

- Decline a guest based on race, color, ethnicity, national origin, religion, sexual orientation, gender identity, or marital status.
- Impose any different terms or conditions based on race, color, ethnicity, national origin, religion, sexual orientation, gender identity, or marital status.
- Post any listing or make any statement that discourages or indicates a preference for or against any guest on account of race, color, ethnicity, national origin, religion, sexual orientation, gender identity, or marital status.

Gender Identity

Airbnb does not assign a gender identity to our users. We consider the gender of an individual to be what they identify and/or designate on their user profile.

Airbnb hosts may not

- Decline to rent to a guest based on gender unless the host shares living spaces (for example, bathroom, kitchen, or common areas) with the guest.
- Impose any different terms or conditions based on gender unless the host shares living spaces with

the guest.

- Post any listing or make any statement that discourages or indicates a preference for or against any guest on account of gender, unless the host shares living spaces with the guest.

Airbnb hosts may

- Make a unit available to guests of the host's gender and not the other, where the host shares living spaces with the guest.

Age and Familial Status

Airbnb hosts may not

- Impose any different terms or conditions or decline a reservation based on the guest's age or familial status, where prohibited by law.

Airbnb hosts may

- Provide factually accurate information about their listing's features (or lack of them) that could make the listing unsafe or unsuitable for guests of a certain age or families with children or infants.
- Note in their listing applicable community restrictions (e.g. senior housing) that prohibit guests under a particular age or families with children or infants.

Disability

Airbnb hosts may not

- Decline a guest based on any actual or perceived disability.
- Impose any different terms or conditions based on the fact that the guest has a disability.
- Substitute their own judgment about whether a unit meets the needs of a guest with a disability for that of the prospective guest.
- Inquire about the existence or severity of a guest's disability, or the means used to accommodate any disability. If, however, a potential guest raises his or her disability, a host may, and should, discuss with the potential guest whether the listing meets the potential guest's needs.
- Post any listing or make any statement that discourages or indicates a preference for or against any guest on account of the fact that the guest has a disability.
- Refuse to communicate with guests through accessible means that are available, including relay operators (for people with hearing impairments) and e-mail (for people with vision impairments using screen readers).
- Refuse to provide reasonable accommodations, including flexibility when guests with disabilities request modest changes in your house rules, such as bringing an assistance animal that is necessary because of the disability, or using an available parking space near the unit. When a guest requests such an accommodation, the host and the guest should engage in a dialogue to explore mutually agreeable ways to ensure the unit meets the guest's needs.

Airbnb hosts may

- Provide factually accurate information about the unit's accessibility features (or lack of them), allowing for guests with disabilities to assess for themselves whether the unit is appropriate to their individual needs.

When guests are turned down.

Hosts should keep in mind that no one likes to be turned down. While a host may have, and articulate, lawful and legitimate reasons for turning down a potential guest, it may cause that member of our community to feel unwelcome or excluded. Hosts should make every effort to be welcoming to guests of all backgrounds. Hosts who demonstrate a pattern of rejecting guests from a protected class (even while articulating legitimate reasons) undermine the strength of our community by making potential guests feel unwelcome, and Airbnb may suspend hosts who have demonstrated such a pattern from the Airbnb platform.

Specific Guidance for Hosts Outside the United States and European Union

Outside of the United States and the European Union, some countries or communities may allow or even require people to make accommodation distinctions based on, for example, marital status, national origin, gender or sexual orientation, in violation of our general nondiscrimination philosophy. In these cases, we do not require hosts to violate local laws, nor to accept guests that could expose the hosts to a real and demonstrable risk of arrest, or physical harm to their persons or property. Hosts who live in such areas should set out any such restriction on their ability to host particular guests in their listing, so that prospective guests are aware of the issue and Airbnb can confirm the necessity for such an action. In communicating any such restrictions, we expect hosts to use clear, factual, non-derogatory terms. Slurs and insults have no place on our platform or in our community.

What happens when a host does not comply with our policies in this area?

If a particular listing contains language contrary to this nondiscrimination policy, the host will be asked to remove the language and affirm his or her understanding and intent to comply with this policy and its underlying principles. Airbnb may also, in its discretion, take steps up to and including suspending the host from the Airbnb platform.

If the host improperly rejects guests on the basis of protected class, or uses language demonstrating that his or her actions were motivated by factors prohibited by this policy, Airbnb will take steps to enforce this policy, up to and including suspending the host from the platform.

As the Airbnb community grows, we will continue to ensure that Airbnb's policies and practices align with our most important goal: To ensure that guests and hosts feel welcome and respected in all of their interactions using the Airbnb platform. The public, our community, and we ourselves, expect no less than this.

Dangerous Speech Project

Susan Benesch is the founder and executive director of the Dangerous Speech Project, a team of experts on how speech leads to violence. We use our research to advise internet companies, governments, and civil society on how to anticipate, minimize, and respond to harmful discourse in ways that prevent violence while also protecting freedom of expression. We warmly welcome critique and feedback on the ideas offered above. To contact us, please visit dangerousspeech.org/contact

Acknowledgments

I am very grateful for comments on an earlier version of this paper and its ideas, from colleagues in academia, at tech companies, and at NGOs, including Chinmayi Arun, Dan Bateyko, Liz Carolan, Connie Chung, Pierre François Docquir, Rob Faris, Tonei Glavinic, David Kaye, Michael Lwin, Colin Maclay, K.S. Park, and Kit Walsh. Tonei Glavinic also contributed invaluable research, ideas, and editing.

This paper was written for and funded by the Israel Democracy Institute and Yad Veshem, and is republished in this format with permission. The full edited volume in which it appears is available at https://perma.cc/E5ZN-C2S4.

Design

CstudioDesign.com