Center for American Progress

Appendix

Recommended Internet Company Corporate Policies and Terms of Service to Reduce Hateful Activities

Contents

Contributors

Lawyers' Committee for Civil Rights Under Law

Color of Change

National Hispanic Media Coalition

Free Press

SPLC Southern Poverty Law Center

Background and definitions

A free and open internet creates immense social value by empowering individual voices, fostering new forms of thought and expression, expanding access to information, and promoting democratic ideals. However, the internet can also be used to engage in hateful activities and to do so at a large scale.

White supremacist and other organizations engaging in hateful activities are using online platforms to organize, raise funds, recruit supporters, and normalize racism, sexism, xenophobia, religious bigotry, and anti-LGBTQIA animus. Online tools have been used to coordinate attacks, including violence, against people of color, immigrants, religious minorities, LGBTQIA people, women, and people with disabilities. This chills the online speech of the targeted groups, curbs democratic participation, and threatens people's safety and freedom in real life.

Because internet tools are largely owned and managed by the private sector and not government, these corporations must be part of the solution to address the promulgation of hateful activities online. This document recommends policies for these corporations to adopt and implement in order to address hateful activities on their platforms. These recommended policies are meant to broadly encompass entities of any corporate form that perform and/or host any of the following services for internet users, whether the entity provides these services directly to the public, through intermediaries, or as an intermediary:

- Social media, video sharing, communications, marketing, or event scheduling/ticketing platforms
- Online advertising, whether directly, as a reseller, or through resellers
- Financial transactions and/or fundraising
- Public chat services or group communications
- Domain names, whether directly, as a reseller, or through resellers
- Websites, blogs, or message boards

Throughout this report and its recommended policies, we refer to these entities as "internet companies," or in the singular as "internet company."

Defining 'hateful activities'

Throughout these recommended policies, we use the term "hateful activities" to mean activities that incite or engage in violence, intimidation, harassment, threats, or defamation targeting an individual or group based on their actual or perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, or disability.

The policies recommended here for internet companies reflect both a commitment to significantly decreasing hateful activities online and a commitment to an open internet. It is important that internet companies respect the free and open nature of the internet by ensuring that all users of online services are treated with respect; that internet companies do not pick winners and losers in the marketplace of ideas; and that internet companies protect the privacy of all users. An appropriate balance reflects the reality that hateful activities threaten individuals, groups, and democratic institutions.

Nothing in these recommended policies is intended to allow for or support a broadband internet access service provider's blocking, throttling, or prioritizing any lawful content. These recommended policies are intended for what are often termed "information service" or "edge" providers and specifically intended for those entities which we have previously described as internet companies in this document.

Furthermore, nothing in these recommended policies is intended to stop internet companies from providing end-to-end encrypted chat services. Nor are these recommended policies intended to encourage internet companies to access or grant others access to the communications provided in such end-to-end encrypted chat services.

Technologies and how people use them are ever-changing. Similarly, as new approaches to ending hateful activities on online services are tried, and results evaluated, some approaches will work better than others. These recommended policies are based on the online tools and information that are available today. Policies and approaches will need to change as technologies, as well as uses, change and as a result of the lessons learned by internet companies and researchers who evaluate data on hateful activities online.

Corporate policy recommendations

Internet companies should adopt and implement the corporate policies described in the next seven sections. A full explanation of internet companies' policies on hateful activities should be easily accessible to users in a language that the users can understand and should especially be available to users in any language with which they use an internet company's services. Similarly, the policies should be easily accessible to any person with a disability who uses a service, consistent with how they use the service.

In the following recommendations, there are both corporate policies that are user-facing—and are recommended to be included in a company's terms of service or acceptable use policies—and those that require changes in how companies manage matters of staff, resources, and governance. The former are described as "model corporate policy/term of service" and the latter as "model corporate policy."

Terms of service and acceptable use policies

Terms of service or acceptable use policies should, at a minimum, make it clear that using the service to engage in hateful activities on the service or to facilitate hateful activities off the service shall be grounds for terminating the service for a user. For instance, while an online payment processor may not be the vehicle through which a group directly engages in hateful activities, the online payment processor should not knowingly allow the group to use its services to fund hateful activities. Not denying services under this example would mean that the online payment processor is financially profiting from hateful activities.

Model corporate policy/term of service

Users may not use these services to engage in hateful activities or use these services to facilitate hateful activities engaged in elsewhere, whether online or offline.

Enforcement

Strong terms of service or acceptable use policies mean very little if they are not effectively enforced. In practice, enforcement varies significantly across internet companies and can vary within an internet company from case to case. This has made it possible for groups and individuals who have engaged in hateful activities online to continue to operate unscathed or to lose access to a service, only to be reinstated later without explanation. Internet companies must have in place an enforcement strategy that recognizes the scope of the problem and reflects a commitment to continuously and significantly diminish hateful activities within their services.

Users and outside organizations should have the ability to flag hateful activities on an internet company's services, but primary responsibility for removing hateful activities from services should sit squarely with the internet company. Enforcement that relies only or primarily on users or outside organizations to flag hateful activities is an insufficient solution that leaves large amounts of hateful activities in place; can be abused; and requires that many users be subjected to hateful activities prior to the internet company removing the violating material, organization, or individual from the services. The insufficiency of a user flagging system alone is especially evident given the sheer volume of online hateful activities and the tendency of such flagger systems to be co-opted by trolls coordinating mass-flagging campaigns to target racial, religious and ethnic minorities, women, and civil rights activists.

Some steps can, however, improve user flagging as one part of an internet company's strategy to stop hateful activities on its services. Under current practices, some internet companies only inform a flagger of actions taken if the internet company agrees with the flagging, while some internet companies do not inform the flagger

of the action taken whether they agree with the flagger or not. These and similar approaches do not fully encourage flaggers to continue flagging, nor do they create a transparent response to hateful activities.

Internet companies should let users who flag what they believe to be hateful activities know what actions the internet company has taken and why, including if the internet company has chosen to take no action. This clarity encourages flagging of hateful activities, increases company accountability, and allows users to know whether their understanding of what hateful activities are is shared by the internet companies and services that they use.

Some internet companies have begun to identify civil and human rights organizations with experience in identifying hateful activities as trusted flaggers, whose flagging is given priority for review and, where appropriate, expedited action to remove violating activities. This approach can encourage civil and human rights organizations to assist internet companies in identifying hateful activities on their services.

In addition to flagging, internet companies should combine technology solutions and human actors to remove hateful activities. Specifically, internet companies should develop computer programs that actively seek to identify hateful activities on their services so that these can be removed. However, automated solutions alone are insufficient, as they may misidentify hateful activities, remove content inappropriately, or miss certain hateful activities. There should also be a sufficiently large, trained team of internet company employees who are cognizant of relevant social, political, and cultural history and context responsible for supplementing automated technologies. Internet companies must ensure that these efforts are tailored to the mission of addressing hateful activities and do not inappropriately invade users' privacy, profile users based solely on their identity or affiliations, or initiate investigations solely based on offensive speech that does not qualify as a hateful activity. The work of both the technological and human efforts should be audited regularly to ensure that they are effectively reducing hateful activities on an internet company's services while also respecting users' speech and privacy.

While internet companies should affirmatively employ technology to reduce the burden of flagging to identify hateful activities, technology solutions are only as effective and accurate as the data and algorithms employed. Given that data can be generated from sources that suffer from intentional or unintentional bias, technology trained on this data, or algorithms reliant on it, can also contain bias. Automated predictions originating from biased data could create unwarranted impacts on groups and individuals based on their characteristics, including characteristics that hateful activities target.[1] For this reason, the evaluation and training policy includes a recommendation that internet companies test automated applications and algorithms routinely for bias in data and results.

Government actors should not be allowed to use internet companies' flagging tools to attempt to remove content they find objectionable as government actors have other means by which to address content concerns. For instance, in the United States there are strong restrictions on what speech can be limited by government and the requirement for due process prior to such limitations. Nothing in these recommended policies should be interpreted to grant additional authority to governments or to allow government extrajudicial influence over internet companies' content.

Model corporate policy

The internet company will do the following: Provide a well-resourced enforcement mechanism that combines technological solutions with staff responsible for reviewing usage of services to ensure that hateful activities are not present. In addition, allow for individuals and organizations—but not government actors—to flag hateful activities, as well as flag groups and individuals engaged in hateful activities. Create a trusted flagger program for vetted, well-established civil and human rights organizations to expedite review of potential hateful activities. Inform flaggers of the results of the company's review of the flagging, including what actions, if any, were taken and why the actions were or were not taken.

Right of appeal

Determining hateful activities can be complicated in some cases. Thus, a user should have the right to appeal any material impairment, suspension, or termination of service, whether that impairment, suspension, or termination of service is in full or in part. This right should allow for an appeal to be made to a separate, neutral decision-maker—someone other than who made the initial determination—with knowledge of the social, political, and cultural history and context within the country or countries from which the user comes and in which people have access to the perceived transgression of the terms of service or acceptable use policy. The user filing the appeal should have the opportunity to present information to advocate for their position.

Model corporate policy/term of service

Any user who is denied service, in whole or in part, for violation of the hateful activities provisions of the terms of service, shall be given the reason for their service denial at the time of denial. The reason shall be provided in a format sufficient for the user to know what specific activities were the reason for denial of service. The user may appeal through an easily identifiable and accessible online process to a higher-level neutral decision-maker with relevant expertise, present evidence supporting their appeal, and be informed of the result of the appeal and its justification in a timely fashion.

Transparency

Both technologies and how people use them change rapidly. To address hateful activities online, it is important to understand what is occurring, what is working, and what is not. To facilitate this understanding, internet companies should be transparent with the actions that they are taking, why they are doing so, and who is affected. These data should be made available online in easily accessible, comprehensive formats that are both human- and machine-readable. This will allow for researchers, scholars, and others to analyze the data to better understand what is happening, make recommendations, and develop best practices.

Model corporate policy/term of service

The internet company will provide to the general public, via easy online access, regularly—meaning at least quarterly throughout the year—and rapidly updated, summary information that describes:

1. The corporate strategy and policies intended to stop groups, state actors, and individuals engaged in hateful activities from using their services
2. The number of hateful activities identified by the company on its services by protected categories—race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, or disability
3. The number of hateful activities identified by the company on its services by type of hateful activity, whether incitement to or engagement in that activity, and whether it was violence, intimidation, harassment, threats, or defamation
4. The number of hateful activities identified by the company on its services broken down by whether this identification was the result of user flagging or some other company action
5. The total number of potentially hateful activities flagged by users, whether the company agreed with the flagging or not
6. The number of potentially hateful activities flagged by users that were found by the company to have been hateful activities under its policies by protected category
7. The type of flagger, including whether the flagger was an individual, organization, and/or trusted flagger
8. The number of times that content was removed as a result of government action or request, broken down by the government entity, actor, or representative making the request, and broken down by whether a legal process was followed and if so, which one[2]
9. How many people have been denied services for hateful activities-oriented violations of terms of service, disaggregated by the quality of denial—whether it was a termination of services in full, denial of services in part, or removal of a specific piece of content
10. Type of victim targeted—group, individual, organization, among others
11. How many users appealed denials of service and the success rates of appeals

Such information shall be published in an aggregate and/or de-identified format consistent with best practices for protecting personally identifiable information of users and shall be made available in human- and machine-readable formats.

Evaluation and training

In their efforts to address hateful activities online, internet companies are testing a variety of techniques that often combine technology-based tests with human assessors to evaluate whether use of their services constitutes hateful activities. This has not always been successful, because the programmers and human assessors may lack expertise on hateful activities for a variety of reasons, including that they are not properly trained or lack an understanding of the cultural, social, and political history and context of the locales, regions, country, or countries which will have access to the content created. To address this, internet companies should hire recognized experts who have a demonstrated expertise on hate, such as peer-reviewed publications and solid academic credentials directly relevant to germane topics, to advise programmers, develop training content, and oversee training of assessors.

Larger internet companies that operate internationally should locate their assessment operations such that cultural, social, and political history and context are consistent with large user populations. For example, outsourcing assessment to contractors in other countries where there is little knowledge of the United States' cultural, social, and political history and context, almost ensures errors in the enforcement of these terms of service.

Internet companies should engage researchers to track the effectiveness of company efforts to respond to hateful activities performed on or facilitated by their services and then use that research to improve company efforts to remove hateful activities. A recent study by researchers at the Georgia Institute of Technology tracking the outcome of banning hate-filled subreddits can be used as a model to track what happens when an internet company does act to address hateful activities.[3]

Model corporate policy

The internet company will establish a team of experts on hateful activities with requisite authority who will train and support programmers and assessors working to enforce anti-hateful activities elements of the terms of service, develop training materials and programs, as well as create a means of tracking the effectiveness of any actions taken to respond to hateful activities. These experts will report to the senior manager charged with overseeing the addressing of hateful activities companywide and will approve all training materials, programs, and assessments.

The internet company will do the following: Routinely test any technology used to identify hateful activities to ensure that such technology is not biased against individuals or groups based on their actual or perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, or disability; make the training materials available to the public for review; locate assessment teams enforcing the hateful activities rules within affected communities to increase understanding of cultural, social, and political history and context.

Governance and authority

Several factors have increased corporate interest in addressing the reality that groups are using their tools to engage in hateful activities. These include the neo-Nazi march in Charlottesville, Virginia; new European rules that embrace large fines for internet companies failing to address criminal violations under their laws;[4] and revelations by large internet companies that foreign governments or aligned entities have engaged in hateful activities on social media platforms in an attempt to create divisions within democracies. Prior to these recent occurrences, organizations working to oppose hateful activities online found that while some internet companies were willing to meet with them, this seldom resulted in meaningful action or organization wide commitment to change.

Elevating the importance of addressing hateful activities within internet companies is essential to significantly limiting the use of internet companies' services to facilitate hateful activities. To achieve this, internet companies should make addressing hateful activities a role for both their board of directors and senior management. Internet companies should also seek outside expertise to give them a reality check on what is working and what is not. This has been done for years by TV networks to gauge their success in addressing front-of-camera diversity issues.

Model corporate policy

The internet company will integrate addressing hateful activities into the corporate structure in three ways:

1. Assign a board committee with responsibility for assessing management efforts to stop hateful activities on their services.
2. Assign a senior manager, with adequate resources and authority, who is a member of the executive team, to oversee addressing hateful activities companywide and name that person publicly.
3. Create a committee of outside advisers with expertise in identifying and tracking hateful activities who will have responsibility for producing an annual report on effectiveness of the steps taken by the company.

State actors, bots, and troll campaigns

Large-scale initiatives to promote hateful activities may originate with countries or other entities that intend to sow discord or to influence the outcomes of elections. It is clear now that this has happened with foreign actor efforts targeting elections in the United States and multiple countries in Europe. There are also now reports of large-scale social media troll campaigns engaging in hateful activities targeting ethnic or religious groups in both African and Asian countries. These coordinated campaigns of hateful activities have occurred using large numbers of bots and/or large teams of human operatives, both of whom present themselves as someone other than who they are.

Removing hateful activities from online services will require dealing directly with these large-scale initiatives. At their core, these initiatives rely on the ability for anonymous, clandestine, and/or delusive actors, whether human or bots, to manipulate services through coordinated action, especially on social media platforms. In addition to propagating hateful activities, this manipulation provides untrue information to internet companies' users and potentially undermines the legitimacy of platforms, including the many valid and valuable purposes for anonymity and privacy-protective services.

Internet companies must stop the inappropriate use of bots and "troll armies" or "web brigades" that manipulate platforms to undertake hateful activities through coordinated campaigns. Different internet companies have different business models. For some internet companies, taking additional steps to ensure that people are who they say they are is consistent with their business model and can be an important step in stopping these hateful activities. For other internet companies, the opportunity for user privacy and anonymity is something they and their users value. Thus, the approaches to stopping bots, troll armies, or web brigades from engaging in hateful activities may be different from company to company. However, a commitment to anonymity cannot be a reason to not address hateful activities. Similarly, a commitment to users disclosing who they are has not in and of itself stopped these kinds of hateful activities on social media platforms.

When not used for hateful activities, online coordinated campaigns involving people can present a unique opportunity to educate the public and build support for social causes. Internet companies' solutions to hateful activities promulgated by bots, troll armies, or web brigades should not hinder opportunities for collective action on their services. Specifically, while internet companies may be able to use automated tools to identify bots engaged in hateful activities, it is important that well-trained human evaluators are part of any review of potential hateful activities undertaken by coordinated campaigns that involve people on a company's services. There are also potential uses of bots, for example for research, that can be beneficial, and nothing in these policies is intended to discourage the use of bots for these purposes.

Ultimately, internet companies must build effective technology and human-resourced efforts to eliminate the use of bots, troll armies, and web brigades to facilitate hateful activities on their services.

Model corporate policy/term of service

The use of bots or teams of people to create or administer coordinated campaigns that engage in hateful activities is prohibited on the service. The internet company will establish and maintain a variety of effective techniques to consistently and aggressively identify and remove the promulgators of such coordinated campaigns from its services. As with other service denials, people who are denied access to services in full or in part have a right to appeal.

Endnotes

1 This refers to the following characteristics described in the definition of hateful activities: "actual or perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, or disability."

2 These model policies call for internet companies to disallow any use of content-flagging tools by government entities, actors, or representatives. In some jurisdictions, there are separate mechanisms that a government may use to attempt to remove content, such as through a court order. When this occurs, internet companies should report these types of government actions and their results.

3 For more information on this study completed by researchers from the Georgia Institute of Technology, Emory University, and the University of Michigan, see Devin Coldewey, "Study finds Reddit's controversial ban of its most toxic subreddits actually worked," TechCrunch, September 11, 2017, available at https://techcrunch.com/2017/09/11/study-finds-reddits-controversial-ban-of-its-most-toxic-subreddits-actually-worked/.

4 BBC, "Germany starts enforcing hate speech law," January 1, 2018, available at https://www.bbc.com/news/technology-42510868.