Recommended Policies for Tech Companies to Curb Hateful Activities Online

Background and Definitions ................................................................. 2
Core Recommendations ...................................................................... 3
Terms of Service and Acceptable-Use Policies ................................... 3
Enforcement ..................................................................................... 3
Right of Appeal ................................................................................ 6
Transparency ................................................................................... 6
Evaluation and Training ................................................................... 7
Inauthentic Behavior and Algorithmic Bias ...................................... 8
Governance and Authority ................................................................. 9
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, spread disinformation and chill free expression through targeted harassment, the encouragement of violence, and the use of threats and intimidation at a large scale. These activities have been proven to incite hate crimes, create an atmosphere of fear and distrust, endanger public health and safety, and chill free speech and civic participation.

Because the private sector largely owns and manages internet tools, these corporations have a responsibility to address the promulgation of hateful and harmful activities online. This document recommends measures for these corporations to implement to address such activities on their platforms. These recommendations are designed for any corporate entities that:

• Provide social media, video sharing, communications, marketing, or event scheduling/ticketing platforms
• Sell advertising online
• Facilitate online financial transactions and/or fundraising
• Provide public chat services or other online group communications
• Build or host websites, blogs or message boards

These recommendations are intended for corporate entities that perform such services for internet users, whether the services are provided directly to the public, through intermediaries or as an intermediary.

*Throughout these recommended policies, we refer to these entities as “Internet Companies” or in the singular as “Internet Company”.*

*Throughout these recommendations we use the term “hateful activities” to describe 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, caste, gender, gender identity, sexual orientation or disability.*

*We recognize that false information campaigns can be created and spread with an intent to incite, harass or otherwise harm an individual or group based on their identity. Internet Companies should address these tactics in their efforts to minimize hateful activities.*

These recommendations reflect both a commitment to significantly decreasing hateful activities and other malignant efforts to sow division or violence online — and a commitment to an open internet. It is important that Internet Companies respect the internet’s free and open nature 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 and civil and human rights of all users. An appropriate balance reflects the reality that hateful activities threaten individuals, groups and democratic institutions.
Nothing in these recommendations is intended to allow for or support a Broadband Internet Access Service provider’s blocking, throttling or prioritizing any lawful content.

In addition, nothing in these core recommendations is intended to stop Internet Companies from providing end-to-end encrypted messaging 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. However, Internet Companies providing end-to-end encryption should take these issues seriously, including by taking steps to consult with experts on possible mitigating measures.

These recommendations are based on the online tools and information that are available today. Policies and approaches will need to change as technologies and uses change. They will also need to shift in response to lessons learned by Internet Companies and researchers who evaluate data on hateful activities online.

**Core Recommendations**

Internet Companies should implement policies that reflect the summary recommendations described in the next seven sections. A full explanation of Internet Companies’ policies on hateful activities should be easily accessible to users in languages that users can understand. At a minimum such policies should be available to users in any language with which they use an Internet Company’s services. Similarly, the policies should be easily accessible to anyone with a disability who uses a service, consistent with how they use the service.

**Terms of Service and Acceptable-Use Policies**

Terms of Service, Acceptable-Use Policies and Community Standards should at a minimum make it clear that they shall be grounds for terminating the service for a user if he or she utilizes the service to engage in hateful activities on the service or to facilitate hateful activities off the service. 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. Failing to deny services once identified under this example would mean that the online payment processor is financially profiting from hateful activities.

**Core Recommendations:** Users may not use these services to engage in hateful activities or use these services to facilitate hateful activities engaged elsewhere.

**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 enforcement strategies that recognize the scope of the problem and reflect a commitment to continuously diminish hateful activities.
within their services. Given the breadth of moderation approaches across Internet Companies, enforcement may vary based on the kinds of services Companies provide. Internet Companies should always make a tailored good-faith effort, based on their resourcing and reach, to enforce their Terms of Service or Acceptable-Use Policies. Companies should also provide regular opportunities for third-party external reviews and publicly report their enforcement practices.

Users and outside organizations should be able to flag hateful activities on an Internet Company’s services regardless of whether the hateful activity or threat is targeted at the individual user. But the Internet Company bears the primary responsibility for removing said activities from its services. Enforcement that relies only or primarily on users or outside organizations to flag hateful activities is an insufficient solution that leaves significant amounts of hateful activities in place. This approach can be abused and means that many users will be subjected to hateful activities prior to the Internet Company removing the violating content, organization or individual from the services. The insufficiency of a user-flagging system 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.

There are steps, however, that can improve user flagging as part of an Internet Company’s strategy to combat hateful activities on its services. Under current practices, some Internet Companies inform a flagger of actions taken only if the Internet Company agrees with the flagging. Other Internet Companies do not inform the flagger of the action taken regardless of whether they agree with the flagger. These and similar approaches do not encourage flaggers to continue flagging, nor do they create a transparent response to hateful activities.

Internet Companies should let users who flag 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, strengthens company accountability, and enables users to know whether the Internet Company shares their understanding of what hateful activities are.

Some Internet Companies have begun to identify civil- and human-rights organizations with experience in identifying hateful activities as “trusted flaggers.” These companies have prioritized flagging from these groups, and, where appropriate, have expedited their flags to remove violating activities. This approach can encourage civil- and human-rights organizations to assist Internet Companies in identifying hateful activities on their services, but this remains insufficient. It is critical that Internet Companies hire staff who are aware of cultural or geographical nuances, with civil- and human-rights expertise, to inform content-policy design and implementation.

In addition to flagging, Internet Companies should combine technology solutions and human actors to remove hateful activities across all languages. Specifically, Internet Companies should develop computer programs that actively seek to identify hateful activities on their services and in all languages in which the Companies operate, so these can be removed.

However, relying primarily on automated solutions is insufficient as they may misidentify hateful activities and may remove content inappropriately or may miss certain hateful activities. There should also be a sufficiently large team of trained Internet Company employees, supported by humane working conditions and sufficient mental-health resources, who are responsible for supplementing automated
technologies across all languages. Such employees should be aware of relevant social, political, linguistic and cultural history and context. 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 based solely on offensive speech that does not qualify as hateful activities.

There should be regular audits of both the technological and human efforts to ensure that they are effectively ending hateful activities on an Internet Company’s services while respecting users’ speech and privacy. The results of these audits must be shared publicly and transparently to ensure that Internet Companies progress toward ending hateful activities.

Government actors should not be allowed to use Internet Companies’ flagging tools to attempt to remove content they find objectionable as government actors can address content concerns via other means. For instance, in the United States there are strong restrictions on what speech the government can limit and due process must be followed prior to the use of such limitations. Nothing in these recommended policies should be interpreted to grant additional authority to government or to allow government extrajudicial influence over Internet Companies’ content.

Many Internet Companies have granted special exemptions to official accounts, government actors and powerful people, allowing them to promote hateful activities, disinformation and other divisive behavior. Instead, these actors should be held to the same standards (if not higher standards) as regular users. There should be no special exemptions that allow the powerful to spread hate with impunity. Many official accounts at various social-media companies have circumvented platform policies despite promoting hateful activities, disinformation and other divisive behavior. Policies should apply equally to all users and must be enforced.

There may be instances where content created by a state actor or official account is newsworthy or can provide others with an opportunity to engage with different perspectives, satire or critiques of hateful activities. But the Internet Company should not give the state actor or official account special treatment.

**Core Recommendation:**

1. 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 any language and or country where the Company does business.

2. Apply terms of service and other moderation and enforcement practices equitably for all users.

3. Allow individuals and organizations — but not government actors — to flag hateful activities, as well as groups and individuals engaged in such activities.

4. Maintain a trusted flagger program for vetted, well-established civil- and human-rights organizations to expedite review of potential hateful activities across languages.

5. 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

Because the issues involved in defining hateful activities can be complicated, 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 a user to make an appeal to a neutral decision-maker within the company — someone who did not make the initial determination. This neutral decision-maker should have knowledge of the social, political, linguistic, and cultural history and context within the country or countries the user comes from. The appealing user should have the opportunity to present information to advocate for their position.

Core Recommendations:

1. 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 and immediately given instructions on how to appeal the denial.

2. The Company should make the reason(s) for the denial clear and should present those reasons in the language the user operates on the platform.

3. 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 learn the result of the appeal and its justification in a timely manner.

Transparency

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 they are taking and explain why they are doing so. Companies should regularly collect data about their practices and the aggregate users impacted by their moderation and enforcement practices. This data should be made regularly available online in easily accessible, comprehensive formats that are both human- and machine-readable. This will allow for independent researchers, scholars and others to analyze Internet Companies’ business models and data to better understand what is happening, make recommendations and develop best practices to mitigate harm of hateful activities.

Core Recommendations: Provide to the general public, via easy online access, and at quarterly intervals, information that describes:

1. the Company’s 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 category (race, color, religion, national origin, ethnicity, immigration status, caste, 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 and whether violence, intimidation, harassment, threats, defamation or illegality);
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 or not the company agreed with the flagging;

6. the number of potentially hateful activities flagged by users that the company found were hateful activities under its policies by protected category;

7. the type of flagger, including whether the flagger was an individual, organization, trusted flagger, or government entity, actor or representative;

8. how many people have been denied services for 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);

9. the type of victim targeted (group, individual, organization, etc.);

10. how many users appealed denials of service and the success rates of appeals; and

11. to the extent it is available, the aggregate demographic information of users who created the user-generated content subject to content moderation. This information should also be broken down by language, region and country.

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 across languages.

**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 is not always successful because the programmers and human assessors may lack expertise on hateful activities for a variety of reasons. In many instances they are not properly trained or lack an understanding of the political, social, linguistic, and cultural history and context of the locales, regions, country or countries that have access to the content created. There is evidence that content moderators suffer poor treatment across the board and non-English language moderators, specifically, have experienced unequal treatment from their employers. Internet Companies should train community moderators on methods for identifying hateful activities. The Companies should also provide mental-health resources for all moderators they employ.

Internet Companies should also hire recognized experts who have demonstrated deep knowledge about hate and disinformation in peer-reviewed publications and via solid academic credentials directly relevant to germane topics. Such experts should advise Company programmers, develop training content and oversee training of assessors equitably across languages.

Internet Companies that operate internationally should take care to base their assessment operations in places where the political, social and cultural history and context are consistent with large user populations. For example, outsourcing assessment to contractors in countries where there is little
knowledge of the United States’ political, social and cultural history and context almost ensures errors in enforcement of anti-hateful activities and elements of 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 across languages. Companies should then use that research to improve their efforts to remove hateful activities.

**Core Recommendations:**

1. Establish a team of experts on hateful activities with requisite authority who can train and support programmers and assessors working to enforce anti-hateful-activities elements of the terms of service, develop training materials and programs, and track the effectiveness of any actions taken to respond to hateful activities across languages and cultural contexts.

2. Make the training materials available to the public for review.

3. Locate assessment teams enforcing hateful activities within impacted communities to increase understanding of political, social and cultural history and context.

4. Provide training for community moderators.

5. Provide humane working conditions, including mental-health support for training staff who come into contact with hateful activities during the course of their evaluation and moderation.

**Inauthentic Behavior and Algorithmic Bias**

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.

This happened with foreign actors targeting elections in the United States in 2016, 2018 and 2020, and in multiple countries across the globe. There have also been large-scale social-media troll campaigns engaging in hateful activities targeting racial, ethnic or religious groups. However, such tactics are not simply foreign interference campaigns. Instances of home-grown efforts abound, including those organized around hateful activities. These coordinated campaigns have been amplified by the algorithms in place at many Internet Companies. As a result, the reach of hateful activities is a profit-making venture for Internet Companies that choose profits over public safety and human rights.

Removal of hateful activities from online services will require dealing directly with these large-scale initiatives and the systemic AI undergirding their spread. At the core, these tactics rely on the ability for anonymous, clandestine and/or deceitful actors to manipulate services through coordinated action, especially on social-media platforms. In addition to propagating hateful activities, this manipulation provides false and dangerous 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 need to stop the inappropriate use of inauthentic behavior — ranging from use of artificial intelligence through deepfakes, bots, “troll armies” or “web brigades” — that manipulate platforms in often coordinated ways to spread hateful activities and disinformation. Different Internet Companies have different business models. For some Internet Companies, taking additional steps to moderate inauthentic behavior is consistent with their business model and can be an important step in
stopping these hateful activities. Other Internet Companies and their users value the opportunity for user privacy and anonymity.

Thus, the approaches to stopping inauthentic behavior may vary from company to company. However, a commitment to anonymity isn’t justification for ignoring 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 they aren’t used for hateful activities, online coordinated campaigns can present a unique opportunity to educate the public and build support for social causes. Internet Companies’ solutions to hateful activities promulgated by inauthentic behavior 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, well-trained and linguistically local human evaluators must be part of any review of potential hateful activities undertaken by coordinated campaigns that involve people on a company’s services.

Ultimately, Internet Companies must build effective technology absent algorithmic bias — and leverage well-trained human resourcing across languages to eliminate amplification of hateful activities on their services.

**Core Recommendations:**

1. The service prohibits the use of inauthentic behavior, including deepfakes, bots or teams of people to create or administer coordinated campaigns that engage in hateful activities and disinformation.

2. The 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.

3. People who are denied access to services in full or in part have a right to appeal.

4. The Internet Company should employ bias-free algorithms that do not amplify or boost engagement of hateful activities and should employ well-trained and linguistically local human moderators to support such efforts.

5. The Company should conduct regular internal audits of the efficacy and potential bias impact of their algorithms, making their findings easily and publicly available.

**Governance and Authority**

Elevating the importance of addressing hateful activities within Internet Companies is essential to significantly limiting the use of Internet Companies’ services for facilitating hateful activities. To achieve this, Internet Companies should make addressing hateful activities a role for both their board of directors and senior management, with civil- and human-rights issue expertise reflected in C-suite executives. Internet Companies should also seek outside expertise to give them a reality check on what is working and what is not across languages. Television networks have done this for years to gauge their success in addressing in front-of-camera diversity issues.
Internet Companies that develop independent structures, such as external boards, to assess the Companies’ activities should give those structures equal access to information across the Companies’ moderation, training and enforcement practices.

**Core Recommendations:** 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 across languages and linguistic groups on their services;
2. assign a member of the executive team, with adequate resources and authority, to oversee addressing hateful activities company-wide and name that person publicly; and
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 across languages.