BAD FACTS MAKE BAD LAW:
HOW PLATFORM CENSORSHIP HAS FAILED
SO FAR AND HOW TO ENSURE THAT THE
RESPONSE TO NEO-NAZIS DOESN’T MAKE
IT WORSE

Cindy Cohn*

CITE AS: 2 GEO. L. TECH. REV. 432 (2018)

INTRODUCTION

From Cloudflare’s headline-making takedown of the Daily Stormer to YouTube’s summer restrictions on LGBTQ content, 2017 was a banner year for platform censorship.¹ Companies—under pressure from lawmakers, shareholders, the press, and some members of the public—ramped up restrictions on speech by adding new rules, adjusting their still-hidden algorithms, and hiring more staff to moderate content. They banned ads from certain sources and removed offensive but legal content, based mostly on complaints by some users against others and algorithms tracking those complaints. Groups, both formal and informal, embarked upon campaigns to convince all platforms, from Facebook to Github to GoDaddy, to drop Neo-Nazi groups;² just as other groups joined together

¹ In this article I use the term “censorship” in the broader sense, referring to when one entity prohibits speech from reaching an audience. Lawyers often use the term to refer only to the situation in which the government prevents speakers from speaking. In my experience, the public does not generally make such distinctions, focusing on the act of prevention, not the entity doing the prevention, so I use the term according to that meaning. See, e.g., ENGLISH OXFORD LIVING DICTIONARIES, https://en.oxforddictionaries.com/definition/censorship [https://perma.cc/X7PC-D685].

² Becky Peterson, Cloudflare is Helping Defend a Neo-Nazi Website from Hackers, Even as Google and GoDaddy Are Distancing Themselves from it, BUS. INSIDER (Aug. 14, 2017, 4:09 PM), http://www.businessinsider.com/cloudflare-defends-nazi-website-from-
to urge censorship in the other direction, with complaints about those who complained about Neo-Nazi and other hateful groups. Many pundits and traditional media voices have urged online platforms to even more aggressively and proactively police the speech of their users.³

But the demands were not limited to just those pro and against Neo-Nazi groups. Censorship demands also impacted people of color and other marginalized groups far beyond that dispute, some as direct targets and others as apparent side effects.⁴ The Black Lives Matter Movement was even included in an FBI report which suggested “black identity extremists” were an emerging kind of terrorist, setting that group up for more takedowns by the platforms that host their speech.⁵

The emotional energy behind this pressure is understandable. In 2017 we saw offline violence by neo-Nazis who partially organized online.⁶ We saw increasing political polarity around the world. We saw ongoing and often increasing harassment of women and people of color online. Those peddling hate seem to be gaining power in the United States and around the world. It is not surprising that these events sparked fierce public calls for action in the online space. Germany even enacted a law that imposes fines of up to fifty million euros on social networks that fail to promptly remove hate speech, fake news, and other undesirable content.⁷

³ See Sean McElwee, The Case for Censoring Hate Speech, HUFFPOST (July 24, 2013, 12:12 PM), https://www.huffingtonpost.com/2013/07/24/hate-speech-online_b_3620270.html [https://perma.cc/L33L-9HR7].
⁴ See infra, note 36–39.
The “platforms should censor more” response to hate online may feel like a good solution to many concerned by rising hate around the world, but in this paper I hope to convince you that the reality is more complicated. It requires addressing some hard questions, first among them: is more censorship effective in stopping hateful acts offline? What are the pitfalls and unintended consequences of this approach? Who actually gets censored under this approach? The truth is that this understandable impulse often leads to bad outcomes, and worse, is vulnerable to manipulation. Even without active manipulation, this censorship does not differentiate well between hateful speech and speech that draws attention to the hatred. Can the current platform censorship processes be made better? Are there even better options than censorship? This paper tries to answer these questions based on the past decade of experience that the Electronic Frontier Foundation (EFF) has had in assisting those facing attempts to censor them online.

Before we begin, though, it is important to pause for two threshold questions. First, much of the current focus is on Internet platforms, even as broadcast media like Fox News on television and Alex Jones and Rush Limbaugh on radio have been promoting hate against people of color and other marginalized groups since long before Facebook or Twitter even existed. They continue to do so. Despite this, the increasing calls for Facebook and Google to censor are not being matched by calls against the hate and division sown by those with longstanding broadcast platforms. These broadcast outlets plainly have an important role in feeding the hate that now spills onto the Internet, one that may be even bigger than the platforms. There is a danger that by focusing on the Internet platforms, we may miss one of the roots of the problem in favor of the flavor of the month.

Second, as my first question notes, the focus on hate speech can blind us to the underlying problem of hate itself. Will chasing this speech out of the public sphere reduce the hateful actions we are seeing in the offline world? Will the people spreading this hate be less successful if they were driven to less public parts of the online world, of which there are many? Or if the calls for hate are limited to broadcast media? Will platform censorship cause the tide of hate to recede? Without answers to these questions, the current efforts aimed to achieve more platform censorship may not be the best use of our resources.

**Executive Summary**

While the threshold questions are important, much of the current focus is on convincing the Internet platforms to censor more and better. Since EFF’s and my personal expertise is also in the online arena, let’s
take a clear-eyed view of this approach. This paper has three parts. First, we look at the now decade-long record of censorship by online platforms. EFF has handled, watched, or participated in many situations arising out of platform censorship over the past decade (and earlier). This has given EFF a wealth of experience in how censorship-based approaches work in practice. One important lesson obtained from this experience is that while a censorship approach might be appropriate for direct hosts of speech like Facebook and Twitter, indirect platforms and services, like Cloudflare’s DDOS protection and upstream web hosts like Hurricane Electric and others, should not be pressed into censorship. They have practical limitations that render this approach overbroad and dangerous. In short, indirect hosts are more likely to make mistakes, more likely to respond to powerful pressure against less powerful targets, less likely to stand up for individual users, and all too often unable to respond precisely, resulting in censorship that is overbroad, often hugely so.

Second, even as to direct hosts of speech, censorship of users must greatly improve from how it has been done in the past. Censorship is a blunt instrument. Wielded by technology alone or by humans without proper oversight, training, and context, it can go horribly awry. Any censorship scheme must include a clear recognition that both user-directed and algorithmically-directed processes (which are generally trained on user submissions) will be vulnerable to both mistake and misuse, including by individuals and organizations seeking to weaponize the scheme to silence their enemies. A platform must include mitigation strategies for mistake, manipulation, and other problems that censorship can cause. The problems are especially acute when powerful entities seek to use censorship to silence less powerful voices. Chief among the mitigation strategies are basic due process protections and expeditious and easy error correction. Finally, direct hosts should be pushed toward non-censorious (or less censorious) ways than platforms to achieve the goal of building an online community that reduces, rather than empowers, hateful speakers.

THE LANDSCAPE

From a legal standpoint, private companies have the power to censor their users. As private entities, the First Amendment protects platforms’ right to restrict the messages their users can convey. As a statutory matter, Section 230 of the Communications Decency Act (47 U.S.C. § 230) protects platforms from most civil and state criminal law

---

liability for most of what their users say.\(^9\) Importantly, Section 230 does not protect against federal criminal liability. On the other side, Section 230 also protects platforms from civil claims by censored users.\(^10\) The goal of Section 230 often gets lost in the current public debates, in which the platforms themselves are often the focus and cutting back on the law is presented as a club against them. Yet, the statute is ultimately aimed at making sure ordinary people can speak online. It does so by ensuring that platform owners can censor to curate their platforms and can be protected against civil and most state criminal liability for the words of their users, as well as against liability from any censored user.\(^11\) Quite simply, Section 230 recognizes the reality that no technology provider would be willing to host user speech, especially at scale, if it was held liable for the criminal and civil claims that could be leveled against its users’ speech.

In practice, all major direct online platforms censor. They set forth rules for their users. Violations of these rules can result in the user’s speech not being published or being taken down after initial publication. Violations can also prompt account suspensions or users being banned entirely. The rules tend to be complex, covering everything from terrorism and hate speech to copyright and impersonation. Most platforms use a version of community reporting to assist in identifying objectionable content, meaning that other users can have a huge voice in whether a person is censored or banned.\(^12\) Every major direct platform (more on indirect ones below) has growing teams that administer these rules and respond to community complaints. Reports indicate that the job of responding to community complaints is highly stressful and has high turnover.\(^13\)

---


\(^10\) § 230(c)(2).

\(^11\) § 230(b).


While the legal situation is clear and the facts on the ground indicate that there is already a significant amount of platform content censorship occurring, there are also long-recognized dangers and harms of a “censor more” approach to combatting hate speech.

I. PLATFORM CENSORSHIP IS NOT NEW; IT HURTS THE LESS POWERFUL AND DOES NOT WORK

Widespread public interest in censorship by intermediaries may be new, but platform censorship is not. Many of the current headline-driven discussions often fail to grapple with the size of the problem or examine the actual record that tech companies have made for themselves so far in this area. Yet, we have well over a decade of evidence documenting how these rules are used and misused.

A. Indirect Intermediaries Should Not Censor Based Upon Content

First, as noted above, not all platforms are equally suited to censor, and even fewer are suited for the aggressive censorship being called for. Internet speech depends on an often-fragile consensus among many systems and operators. On the one hand, there are direct hosts, such as social networks like Facebook and Twitter; hosted websites like Github, which hosts many coding projects; or Megafilter, a community weblog. On the other hand, there are indirect intermediaries, also called “upstream intermediaries,” such as web hosting services, DNS providers, Internet Service Providers (ISPs), domain name registrars, and services, such as Cloudflare’s DDOS protection. While all of these hosts are necessary for speech to reach its audience, they are not all the same when they are asked to censor.

Indirect intermediaries are generally far removed from the speech itself. Domain name hosts and DNS providers assign IP addresses and names for websites. ISPs often host entire, complex websites. Cloudflare and similar services allow websites to be seen online in the face of certain kinds of attacks. None of these have as direct a relationship with online speakers as platforms like Facebook or Twitter.

While each indirect intermediary provides a slightly different service, they are similar in that they are often unable to remove only a single post, and instead, they often remove an entire website, domain, or...


14 See generally Free Speech Only as Strong as the Weakest Link, ELEC. FRONTIER FOUND., https://www.eff.org/free-speech-weak-link [https://perma.cc/96PA-RXRG].
worse, a set of domains. They also generally have only limited technical interactions with their customers. That is why EFF has called these parts of the Internet “free speech’s weakest links.”\(^{15}\) Censorship here has much larger implications than that aimed at direct hosts, as seen in the following examples of takedowns EFF has handled that were aimed at “indirect” intermediaries:

- **Hosting Service:** The U.S. Chamber of Commerce sent a complaint about a parody website created by activist group, The Yes Men, to its hosting service, May First/People Link, and to that service’s upstream ISP, Hurricane Electric. When the hosting service May First/People Link resisted demands to remove the parody site, Hurricane Electric shut down MayFirst/PeopleLink’s connection entirely, temporarily taking offline hundreds of "innocent bystander" websites as collateral damage due to its fear of the Chamber.\(^{16}\)

- **Domain Name Registrar:** Unhappy with a single document published on Cryptome.org, Microsoft sent complaints to Cryptome’s domain name registrar and web hosting provider, Network Solutions. As a result, hosting provider Network Solutions removed the entire Cryptome website—full of legal content—because Network Solutions was not technically capable of targeting and removing the single document. Public outcry ultimately forced Microsoft to retract its takedown request.\(^{17}\)

- **Domain Name Registrar:** In 2008 Swiss bank Julius Baer pressured Wikileaks’ domain name registrar to lock Wikileaks’ domain name after the organization posted documents demonstrating financial wrongdoing. The bank then obtained a court ruling confirming the seizure of the Wikileaks.org domain name. In response to legal briefs by EFF and others

---

\(^{15}\) Id.


 objected to this tactic, the district court dissolved the order, which led Julius Baer to dismiss its case.\textsuperscript{18}

- **Domain Host**: Threats to the domain host of a critic of South African diamond conglomerate, De Beers, resulted in the temporary takedown of an entire website containing a *New York Times* spoof. The diamond company took issue with a critical fake ad. The entire website was removed because the domain host could not remove a single ad.\textsuperscript{19}

- **Website Host**: Media giant ABC sent a cease and desist letter on behalf of KSFO-AM radio in San Francisco to the webhost of the blog www.spockosbrain.com, after that site criticized offensive and violent rhetoric on the radio station aimed at Congresswoman Nancy Pelosi and then-Senator Barak Obama. Again, the entire website was threatened even though the complaint was only on a single blog post.\textsuperscript{20}

- **Internet Service Provider**: Shell Oil sent a takedown notice to the ISP of activist group Oil Change International after it launched a campaign aimed at Shell’s sponsorship of New Orleans Jazz Fest. The ISP removed the entire site, despite concerns about only a few statements on the pages, but the notice abruptly halted the entire campaign.\textsuperscript{21}

It is no coincidence that these complaints were made by large companies against smaller entities. Large companies have the time, money, and lawyers to pressure intermediaries to do their bidding—something smaller communities rarely have. Especially with indirect


censorship, the relative power of those facing censorship plays a key role in who gets censored and who does not.\textsuperscript{22}

When the horrific events in Charlottesville resulted in first Network Solutions, then GoDaddy, then Google dropping the neo-Nazi Daily Stormer website, many cheered and urged for more of the same. But the examples above, involving multiple instances of overbroad censorship triggered by the powerful and used to silence the less powerful should give us pause before embracing this blunt tool.\textsuperscript{23}

By opening up this avenue for censorship, indirect intermediary Cloudflare found itself on the receiving end of a large number of invalid complaints and at least one lawsuit.\textsuperscript{24} In February 2018, Cloudflare CEO Matthew Prince reported that Cloudflare received over 7,000 requests to take down websites—all citing the Daily Stormer case.\textsuperscript{25} “It spanned the political spectrum,” Prince said.\textsuperscript{26} “People wanted us to take down other neo-Nazi sites, as well as extreme left wing sites. There was even a cooking blog that somebody claimed was incredibly offensive. We looked at it and it’s just a cooking blog. To this day we have no idea why—our internal theory is that the recipes are just so terrible.”\textsuperscript{27}

As the examples demonstrate, indirect intermediaries have little incentive and often lack the necessary context or other information to review a takedown demand, especially when a powerful company or person is objecting to material hosted by a single user. They have even less incentive to set up a process for takedowns that includes due process like investigation, receiving input from all sides or correcting errors. Generally, they have many customers who, at most, pay only small amounts of money for the services received. Time spent supporting a single user can quickly overcome the amount charged for the service, even if the indirect intermediary is not being asked to stand up against a

\textsuperscript{22}While these examples all had relatively happy endings, thanks in part to EFF’s involvement, that should not eliminate concerns about the risks of the strategy of seeking censorship by indirect intermediaries.

\textsuperscript{23}Upstream Providers, ELEC. FRONTIER FOUND., https://www.eff.org/free-speech-weak-link/#upstream [https://perma.cc/C3NH-CZJJ].


\textsuperscript{26}Id.

\textsuperscript{27}Id.
powerful objector. Yet as seen with Cloudflare, if indirect intermediaries open themselves up to censorship demands, the demands will come.

One additional risk was laid bare during the debates over SOPA\(^{28}\) and PIPA,\(^{29}\) two failed bills that sought to use the domain name system and other upstream intermediaries in the service of copyright enforcement. One suggested technical strategy for avoiding upstream censorship was to “fork” the DNS system, which results in two domain naming systems that will not talk, or talk well, to each other.\(^{30}\) While the proposals were defeated in Congress eliminating the need to consider this option, it remains technically available. Ultimately, attempts to use key indirect intermediaries to limit speech risks shattering the consensus that keeps the Internet working as a whole.

For these reasons, EFF continues to urge indirect hosts to resist calls to serve as censors.\(^{31}\)

### B. Censorship by Direct Hosts

What about direct hosts like Facebook and Twitter? A review of onlinecensorship.org, EFF’s joint venture project tracking social media takedowns, can give some perspective.\(^{32}\) Despite the current exhortations to do more, direct hosts have not to date been conservative in their attempts to police speech on their platforms.\(^{33}\) Facebook reported that it deletes 288,000 hate-speech posts alone in a single month.\(^{34}\) In 2016, Twitter boasted that it deleted 135,000 “terrorism-related” accounts from their service.\(^{35}\)

---

31 Cohn, *supra* note 21.
33 See id.
How is that censorship doing? The results are not encouraging. Private censorship by direct hosts is a problem not just because of the slippery slope from “good” censorship to “bad” censorship but because we are already far down that slope. Indeed, within three days of the launch of the new German law that created liability for failing to take down terrorist speech, Twitter blocked a German satirical magazine’s Twitter account after it parodied anti-Muslim comments.\(^{36}\)

Not only does censorship remove some useful speech, as with indirect intermediaries, marginalized groups are often the first to be impacted by private censorship by direct intermediaries.\(^{37}\) Some examples are:

- Prohibitions on hate speech have been used to:
  - Shut down conversations among women of color about the harassment they receive online.\(^{38}\)
  - Remove a video that showed a white male calling children the N-word.\(^{39}\)
  - Suspend an activist and author called “antifa’s most prominent Jew,” after complaints from alt-right activists.\(^{40}\)
  - Ban lesbians from referring to themselves as “dykes” due to hate speech policies.\(^{41}\)
  - Ban Black Lives Matter activist for reposting a racist message he received.\(^{42}\)

---


39 Tracy Jan & Elizabeth Dwoskin, *supra* note 34.

40 Sam Kestenbaum, ‘Antifa’s Most Prominent Jew’ Booted from Twitter, FASTFORWARD (Jan. 9, 2017), https://forward.com/fast-forward/374276/antifa-s-most-prominent-jew-booted-from-twitter/ [https://perma.cc/C6Z5-AVW3]. Activist Sieradsky alleges that he was removed from Twitter at the urging of Nazis; although the reporter states that his suspension may have been for other reasons.

- Remove a historical World War II video showing the demolition of Nazi symbols.43
- Ban Pepe the frog, an alt-right cartoon used in racist memes, while allowing swastikas as “the display [of] hate symbols for political messaging.”44

- Rules against harassment employed to shut down the account of a prominent Egyptian anti-torture activist.45

- Rules requiring real names have lead to domestic violence victims being found by their abusers.46

- Bans on nudity have been used to:
  - Censor women who share childbirth images in private groups.
  - Take down breastfeeding photos.47
  - Remove a picture of a statue of a nude sea god.48
  - Repeatedly take down a famous Vietnam war photo.49

---

49 Alice Ross & Julia Carrie Wong, *Facebook Deletes Norwegian PM’s Post as ‘Napalm Girl’ Row Escalates*, GUARDIAN (Sept. 9, 2016, 8:29 AM),
• Regulations on violent content have:
  ▪ Removed documentation of police brutality in the U.S. and abroad.\(^50\)
  ▪ Removed documentation of abuses during the Syrian war.\(^51\)
  ▪ Removed human rights abuses suffered by the Rohyinga.\(^52\)
  ▪ Removed pictures and reports of military violence against the Muslim minority in the province of Kashmir.\(^53\)
  ▪ Removed video showing police arresting Dakota pipeline protesters.\(^54\)

This small sampling of content removal censorship demonstrates that platform censorship strategies, especially those based on community complaints, regularly remove things that they should not. More troubling, they also show that these schemes often backfire by preventing community awareness of human rights abuses, the impact of hate speech, and violent behavior of all kinds.

Moreover, as detailed below, the appeal processes provided by social media companies are not sufficient. While complete data is not publicly available, of the 230 reports received by OnlineCensorship.org between March and October 2016, 103 or 44.7% attempted to appeal. Of
those appealing, at least twenty-eight, or 27% said they did not get any response whatsoever from the platform.  

C. Danger to Anonymous Speech

In addition to taking down valuable speech, the pressure to censor leads to attacks on anonymous and pseudonymous speech. Anonymity and pseudonymity have played important roles throughout history, including, most famously, the conversations between the founders of the United States. Online anonymity allows users to explore controversial ideas and connect with people’s concerns without exposing themselves unnecessarily to harassment and stigma. It enables dissidents in oppressive regimes to tell their stories with less fear of retribution.

Study after study has debunked the idea that forcibly identifying speakers is an effective strategy against those who spread bad information online. Counter-terrorism experts tell us that “[c]ensorship has never been an effective method of achieving security, and shuttering websites and suppressing online content will be as unhelpful as smashing printing presses.”

Yet, despite this, current proposals from private companies to increase censorship often undermine online anonymity. For example, Twitter’s recent ban on advertisements from Russia Today and Sputnik relies on the notion that the company will be better at identifying accounts controlled by Russia than Russia will be at disguising accounts to promote

55 See, e.g., Jessica Anderson et al., Censorship in Context, ONLINECENSORSHIP.ORG (Nov. 2016), https://s3-us-west-1.amazonaws.com/onlinecensorship/posts/pdfs/000/000/088/original/Censorship_in_Context_November_2016.pdf [https://perma.cc/L765-SFGC] (Note that EFF has been involved in both of these projects, along with other organizations.).
its content. To be effective, Twitter may have to adopt new policies to identify and attribute anonymous accounts, undermining both speech and user privacy. Given the problems with attribution, Twitter will likely be called to ban anyone from promoting a link to suspected Russian government content.

What will we get in exchange for giving up our ability to speak online anonymously? Very little. We already have a very clear, very public demonstration of this: Facebook. For many years, Facebook required individuals to use their real name (and continues to require them to use a variant of it). Yet people are not nicer to each other on Facebook—hateful speech, including neo-Nazi speech, is alive and well on that platform. Nor did the policy stop Russian agents from gaming the rules or stop the use of bots and other mechanisms to create the appearance of more people in favor of hateful ideas.

Yet, at the same time, the real names policy has undermined innocent people who need anonymity—including drag performers, LGBTQ people, Native Americans, survivors of domestic and sexual violence, political dissidents, sex workers, therapists, and doctors. Anonymity is often the greatest shield that less powerful groups have against powerful governments or corporate forces, yet the “censor-more” based pressure to eliminate anonymity continues.

II. GOVERNMENT-INDUCED “PRIVATE” CENSORSHIP, A.K.A. SHADOW REGULATION

While governmental censorship is generally limited, and, at least in the U.S. includes strict rules and protections to try to stop it from going awry, “private” censorship by the platforms themselves does not. Yet not all “private” censorship by platforms is truly private. Over the past decade a secretive web of backroom agreements has emerged between companies that seeks to control our behavior online, often driven by governments as a shortcut and an alternative to regulation. One example among many: under pressure from the United Kingdom Intellectual Property Office, indirect intermediary search engines agreed last year to a

---


"Voluntary Code of Practice" that requires them to take additional steps to remove links to allegedly unlawful content. At the same time, domain name registrars are also under pressure to participate in copyright enforcement, including “voluntarily” suspending domain names. Similarly, in 2016, the European Commission struck a deal with the major platforms to encourage them to take down “illegal” speech under their terms of service. The effect of this approach is to circumvent the role of judges and courts in determining whether speech is, in fact, illegal.

Whether aimed at direct or indirect intermediaries, shadow regulation is dangerous and undemocratic. Government regulation should take place in the sunshine, with the participation of the various interests and the checks and balances of governmental systems, rather than the unilateral enforcement of contractual terms by private companies.

III. NECESSARY PROTECTIONS FOR CENSORSHIP SCHEMES: DUE PROCESS PROCEDURES

Despite the concerns raised by increasing censorship and the difficulties in doing it well, pressure on the direct hosting companies is likely to continue. As noted above, the lack of due process impacts users. It may be, however, that by increasing investment in the process, including increasing the procedural protections for speakers as well as the contextual sensitivity of reviewers, direct platforms can reduce the amount of hateful and harmful speech on their platforms without a significant increase in censorship of valuable speech. What would better procedures look like?

One high-level place to start could be the due process portion of the Manila Principles, a set of best practices and policies for intermediary liability. More detailed recommendations are gathered in the reports of Onlinecensorship.org, which has been collecting examples of private censorship by social networks. For purposes of this short paper, however, the Manila Principles provide a basic framework.

The Manila Principles were created by an international group of organizations and have been endorsed by over one hundred organizations

---


64 Mitch Stoltz, MPAA May Like Donuts, But They Shouldn’t Be the (Copyright) Police, ELEC. FRONTIER FOUND. (Feb. 10, 2016), https://www.eff.org/deeplinks/2016/02/mpaa-may-donuts-they-shouldnt-be-copyright-police [https://perma.cc/CUS5-HR38].

from fifty-seven countries. The principles have been referenced in reports by United Nations rapporteurs and the Organization for Security and Co-operation in Europe (OSCE), along with the Global Commission on Internet Governance's One Internet report. The Manila Principles include a robust list of due process protections for takedowns at the behest of governments, which, slightly adjusted, can serve as an initial blueprint for platforms engaging in private censorship.

- **Publish Content Restriction Policies Online.** Intermediaries should publish their content restriction policies online in clear language and accessible formats with plentiful examples. They should keep users updated as the policies evolve and notify users of changes when applicable.
  - This should include messaging on the mechanisms used to evaluate content at multiple points of communication across the site and app, including a centralized location for policy information (as YouTube already does with its Policy Center), as well as in-app notifications of content takedowns and “report content” sections.

- **Content Restriction Request Procedures.** Complaints to a platform that a user has violated company policies must, at the minimum, contain the following:
  - The reasons why the content at issue is in breach of the intermediary’s content restriction policies, including identifying the specific policy breached.
  - The Internet identifier and description of the alleged violation of the content restriction policies.
  - Contact details of the issuing party or their agent.

---

• A declaration of good faith that the information provided is accurate.

• **Forward Complaint Material.** With the exception of the identity of the complainant, all complaint material should be forwarded to the accused user. This must include the specific material flagged and the specific policy that was violated. It must also include whether the content was taken down because it was flagged by another user or by another reporting mechanism.

• **Clear and Accessible Explanation of User Rights.** When forwarding the request, the intermediary must provide a clear and accessible explanation of the user content provider’s rights, including a description of and any necessary forms or links for the available counter-notice or appeal mechanisms.

• **Right to be Heard.** Before any content is restricted, the accused user must be provided an effective right to be heard except in exceptional circumstances, in which case an ex post facto review of the order and its implementation must take place as soon as practicable.

• **Right of Appeal.** Any takedown process must provide the accused user the right of appeal against content restriction decisions.

• **Transparency Reports.** Intermediaries should publish transparency reports providing specific information about all content restrictions taken by the intermediary.

• **Notice of Removed Content.** Where content has been restricted, the intermediary must display a clear notice to others seeking access that explains which content has been restricted and the reasons for that restriction.

• **Personally Identifiable Information.** An intermediary should not disclose personally identifiable information about an accused user or an accuser without either an order by a judicial authority or knowing, informed, and specific consent.
• **Human Rights.** When drafting and enforcing their content restriction policies, intermediaries should respect international human rights laws and principles.

IV. **NON-CENSORSHIP OPTIONS: WHAT CAN PLATFORMS DO?**

While due process protections can assist, they cannot entirely cure those problems created by censorship schemes. This is especially the case when well-funded or powerful complainants are aiming at less powerful targets. Thus, it is important to look beyond censorship as a strategy. What non-censorship options are available to platforms?

A. **Meaningful Transparency**

Over the years, groups like EFF have pushed companies to be more transparent about the speech that they take down, particularly when it is at the behest of governments. Transparency reports are helpful, but they are insufficient. When it comes to decisions about acceptable speech, or what kinds of information or advertisements to show us, companies remain largely opaque. The problem is complicated by the algorithmic decision-making that plays a role in the priority and placement of user content.

Facebook, Google, and others should allow truly independent researchers—with no bottom line or corporate interest—access to work with, black box test, and audit their systems. Users should be told when bots are flooding a network or impacting a newsfeed or other algorithmically-driven decision making. Users should also be provided with better tools to avoid seeing these kinds of messages and be able to combine their forces to share knowledge about those messages and messengers. Meaningful transparency also means allowing users to see what types of content are taken down, what is shown in their feed and why. It means being open and straightforward with users about how their data is collected and used.

B. **Empower Users with Much Better Platform Tools**

Users should be in charge of their own online experience. Platforms should continue to build tools such as smart filters that let users choose the content they wish to see. This approach has been criticized for

---

furthering “information bubbles,” but those problems are less worrisome when users are in charge and informed than when, as now, self-interested companies are making these decisions for users with a goal of keeping them on the platform or ensuring that they see certain ads.

Facebook already allows users to choose what kinds of ads they want to see—a similar system should be put in place for content. Facebook’s machine learning models can recognize the content of photos, so users should be able to choose an option for "no nudity" rather than Facebook banning it wholesale. (The company could still check that box by default in countries where it is illegal.) This system should also include tools that let users make those decisions in real time while they are seeing the images, rather than having to find a buried interface.

Additionally, users must be empowered to set limitations on how long and in what ways data they provide, as well as data that is collected based upon their activities on the platform, can be kept and used.

C. Empower Users with Third Party Tools

Big platform companies are not the only place where good ideas can grow. Right now, the larger platforms limit the ability of third parties to offer alternative experiences on the platforms by using closed Application Programming Interfaces (APIs), blocking scraping and limiting interoperability. They enforce their power to limit innovation on the platform through a host of laws, including the Computer Fraud and Abuse Act, the Digital Millennium Copyright Act, and anti-competitive clauses in terms of service. Larger platforms like Facebook, Twitter, and YouTube should facilitate user empowerment by opening their APIs to competing services and by allowing scraping, ensuring interoperability with third party products and forking of services.

D. Forward Consent

Community guidelines and policing are touted as a way to protect online civility but are often used to take down a wide variety of speech. The targets of reporting often have no idea what rule they have violated,

---

73 See, e.g., Terms of Abuse, ELEC. FRONTIER FOUND., https://www.eff.org/issues/terms-of-abuse [https://perma.cc/UZ2L-U479].
since companies often fail to provide adequate notice. The first step is requiring companies to provide real notice to targets, including the specific post or language objected to, as noted above in the due process section. A second step is for platforms to have users affirmatively accept the community guidelines point-by-point and accept them again each time they change.

E. Judicious Filters by Platforms

Platforms can and should use smart filters to better identify content that violates community standards for human review. Filters can also help the company to recognize when their user flagging systems are being gamed either to avoid detection of violative content or by those seeking to get the platform to censor others. However, caution is necessary here. Filtering technologies should not be used to automatically take down speech. The default for online speech should always be to keep it online until a human has reviewed it; although some exceptions to even that rule may be appropriate. For example, where a file is an exact match of a file already found to be infringing, where no effort was made to counter notice.

F. Competition amongst Platforms and User Choice

Ultimately, users also need to be able to leave when a platform is not serving them. Real data portability is key here, and this will require companies to agree to standards for how social graph data is stored and transferred. Fostering competition in this space, while being mindful of privacy risks, could be one of the most powerful incentives for companies to protect users against bad actors on their platform. Pressure on companies to allow full interoperability and data portability could lead to a race to the top for platforms.

CONCLUSION

These are just some starting ideas. While the rise of Neo-Nazis and other hateful speech is rightly troubling, in the rush to respond we should not ignore the very deep problems in pushing a larger censorship obligation onto online platforms. Instead, we should put more work into supporting due process, building in error recovery systems, and empowering users with finding non-censorship ways to combat hate online more directly.