EXTREMIST SPEECH, COMPELLED CONFORMITY, AND CENSORSHIP CREEP

*Danielle Keats Citron**

Silicon Valley has long been viewed as a full-throated champion of First Amendment values. The dominant online platforms, however, have recently adopted speech policies and processes that depart from the U.S. model. In an agreement with the European Commission, the dominant tech companies have pledged to respond to reports of hate speech within twenty-four hours, a hasty process that may trade valuable expression for speedy results. Plans have been announced for an industry database that will allow the same companies to share hashed images of banned extremist content for review and removal elsewhere.

These changes are less the result of voluntary market choices than of a bowing to governmental pressure. Companies' policies about extremist content have been altered to stave off threatened European regulation. Far more than illegal hate speech or violent terrorist imagery is in EU lawmakers' sights, so too is online radicalization and "fake news." Newsworthy content and political criticism may end up being removed along with terrorist beheading videos, "kill lists" of U.S. servicemen, and instructions on how to bomb houses of worship.

The impact of extralegal coercion will be far reaching. Unlike national laws that are limited by geographic borders, terms-of-service agreements apply to platforms' services on a global scale. Whereas local courts can order platforms only to block material viewed in their jurisdictions, a blacklist database raises the risk of global censorship. Companies should counter the serious

© 2018 Danielle Keats Citron. Individuals and nonprofit institutions may reproduce and distribute copies of this Article in any format at or below cost, for educational purposes, so long as each copy identifies the author, provides a citation to the *Notre Dame Law Review*, and includes this provision in the copyright notice.

* Morton & Sophia Macht Professor of Law, University of Maryland Francis King Carey School of Law; Affiliate, Stanford Center on Internet & Society, Yale Information Society Project. I am grateful to Leslie Henry, Sarah Hoyle, Margot Kaminski, Kate Klonick, Emma Llansó, and James Weinstein for closely reading drafts, and to Tabatha Abu El-Haj, Jack Balkin, Susan Brison, Ryan Calo, Chapin Cimino, Will Duffield, Ed Felten, James Fleming, Brett Frischmann, Mike Godwin, Abner Greene, Anil Kalhan, Daphne Keller, Paula Kift, Michael Nelson, Megan Phelps-Roper, Neil Richards, Flemming Rose, Marc Rotenberg, John Samples, Alexander Tsesis, and Benjamin Wittes for advice. Participants in the *Fordham Law Review*'s "Terrorist Incitement on the Internet" symposium, Twitter's Trust and Safety Summit, Drexel Law School's Faculty workshop, the Cato Institute's "The Future of the First Amendment" conference, and the Yale Information Society Project's Lunchtime Talk Series provided helpful feedback. Thanks to Emily Bryant and Ian Königsdörffer for superb research assistance. I owe a debt of gratitude to Susan McCarty for her expert editing, Frank Lancaster for assisting in all things, and Dean Donald Tobin for supporting this research. Special thanks to the editors of the *Notre Dame Law Review*, especially Shannon Lewry, for their insight, help, and patience.

potential for censorship creep with definitional clarity, robust accountability, detailed transparency, and ombudsman oversight.

INTRODUCTION

In 2008, U.S. Senator Joseph Lieberman squared off with internet companies and lost. The dispute concerned the Senator's demand that platforms remove hundreds of Al-Qaeda training videos.[1] Senator Lieberman argued that by keeping up the videos, tech companies were complicit in terrorist recruitment.[2]

Google's YouTube held fast in defense of users' right to express unpopular viewpoints.[3] As Jeffrey Rosen wrote at the time, Google's Nicole Wong and her colleagues worked "impressively to put the company's long-term commitment to free expression above its short-term financial interests."[4] Ignoring the Senator's demands was a safe strategy: any effort to proscribe extremist expression would likely fail given the First Amendment's hostility to viewpoint-based regulations.[5]

American free speech values guided policy decisions in Silicon Valley long after the showdown with Senator Lieberman.[6] Social media companies routinely looked to First Amendment doctrine in crafting speech policies.[7]

---

1 *Lieberman to YouTube: Remove al Qaeda Videos*, CNN (May 20, 2008), http://edition.cnn.com/2008/POLITICS/05/20/youtube.lieberman/.

2 Timothy B. Lee, *YouTube Rebuffs Senator's Demands to Remove Islamist Videos*, ARS TECHNICA (May 20, 2008), https://arstechnica.com/tech-policy/2008/05/youtube-rebuffs-senatorss-demands-for-removal-of-islamist-videos/.

3 *Id.* Ultimately, YouTube took down eighty videos under its terms of service, but many others were left up—to the Senator's dismay. Editorial, *Joe Lieberman, Would-Be Censor*, N.Y. TIMES (May 25, 2008), http://www.nytimes.com/2008/05/25/opinion/25sun1.html.

4 Jeffrey Rosen, *Google's Gatekeepers*, N.Y. TIMES MAG. (Nov. 30, 2008), http://www.nytimes.com/2008/11/30/magazine/30google-t.html.

5 Calls for violence or political disruption generally enjoy First Amendment protection—any regulation of such speech would have to overcome the crucible of strict scrutiny. Alexander Tsesis, Essay, *Terrorist Speech on Social Media*, 70 VAND. L. REV. 651 (2017) (arguing that certain forms of terrorist speech can be regulated without running afoul of free speech norms). Incitement to violence can, however, be regulated in the narrow circumstance that violence is intended, likely, and imminent. *See* Brandenburg v. Ohio, 395 U.S. 444, 447–49 (1969) (per curiam).

6 DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 232 (2014); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. (forthcoming 2018) (manuscript at 27–28).

7 Klonick, *supra* note 6, at 26–28. Tech companies have enormous freedom to choose whether and how to address extremist expression. As private actors, online platforms operate free from First Amendment concerns. Gitlow v. New York, 268 U.S. 652, 666 (1925); Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. REV. 1435, 1439 (2011). Under section 230 of the Communications Decency Act, tech companies enjoy broad immunity from liability related to user-generated content. Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401 (2017).

Twitter, an exemplar of this ethos, was aptly known as "the free speech wing of the free speech party."[8]

From the start, tech companies' commitment to free expression admitted some exceptions.[9] Terms of service and community guidelines banned child pornography, spam, phishing, fraud, impersonation, and copyright violations.[10] Threats, cyber stalking, nonconsensual pornography, and hate speech were prohibited after extended discussions with advocacy groups.[11] The goal was to strike an appropriate balance between free expression and abuse prevention while preserving platforms' market share.[12]

More recently, social media companies have revised their speech policies concerning extremist and hateful expression. Unlike previous changes, however, these revisions were not the result of market forces. They were not made to accommodate the wishes of advertisers and advocates.[13] Instead, they were adopted to stave off threatened European regulation. After terrorist attacks in Paris and Brussels in late 2015, European regulators excoriated tech companies for failing to combat terrorist recruitment on their platforms.[14] Their message was clear: online platforms would face onerous civil

8 Josh Halliday, *Twitter's Tony Wang: "We Are the Free Speech Wing of the Free Speech Party,"* GUARDIAN (Mar. 22, 2012), https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

9 Citron, *supra* note 6, at 232.

10 Twitter largely embraced this approach until 2016. *See, e.g.*, *The Twitter Rules*, TWITTER, https://support.twitter.com/articles/18311# (last visited Nov. 4, 2017).

11 Citron, *supra* note 6, at 229–30; Jessica Guynn, *Twitter to Police Abuse in Major Shift*, USA TODAY (Mar. 1, 2017), http://www.usatoday.com/story/tech/news/2017/03/01/twitter-to-police-abuse-in-major-shift/98559482/. For the past eight years, I have been advising social media companies on speech and safety issues. I am a member of Twitter's Trust and Safety Council and part of a small group advising Facebook on its nonconsensual pornography policies. I do not receive compensation for my work with Twitter and Facebook (or any other company that I advise). I serve as an adviser to the Cyber Civil Rights Initiative, an organization whose work has been instrumental in changing companies' policies related to nonconsensual pornography, see Mary Anne Franks, *"Revenge Porn" Reform: A View from the Front Lines*, 69 FLA. L. REV. (forthcoming 2018), as well as a member of the newly formed Advisory Board for the Anti-Defamation League's Center on Technology and Society and the Anti-Cyber Hate Working Group established by the Anti-Defamation League. Although this Article is informed by that work, it is not based on any confidential information shared with me.

12 Mathew Ingram, *Twitter Tries to Find a Balance Between Promoting Free Speech and Curbing Abuse*, FORTUNE (July 20, 2016), http://fortune.com/2016/07/20/twitter-bans-troll-milo-yiannopolous/.

13 Sapna Maheshwari & Daisuke Wakabayashi, *AT&T and Johnson & Johnson Pull Ads from YouTube*, N.Y. TIMES (Mar. 22, 2017), https://www.nytimes.com/2017/03/22/business/atampt-and-johnson-amp-johnson-pull-ads-from-youtube-amid-hate-speech-concerns.html.

14 Liat Clark, *Facebook and Twitter Must Tackle Hate Speech or Face New Laws*, WIRED UK (Dec. 5, 2016), http://www.wired.co.uk/article/us-tech-giants-must-tackle-hate-speech-or-face-legal-action.

and criminal penalties unless their policies and processes resulted in the rapid removal of extremist speech.[15]

Tech companies accommodated these demands because regulation of extremist speech was a real possibility. Unlike in the United States, in the European Union, there isn't a heavy presumption against speech restrictions.[16] On May 31, 2016, Facebook, Microsoft, Twitter, and YouTube entered into an agreement with the European Commission to remove "hateful" speech within twenty-four hours if appropriate under terms of service.[17] Six months later, the same companies announced plans for a shared database of banned extremist content for review and removal elsewhere.[18]

Nearly a decade later, European lawmakers accomplished what Senator Lieberman could not.[19] By insisting upon changes to platforms' speech rules and practices, EU regulators have exerted their will across the globe. Unlike national laws that apply only within a country's borders, terms of service apply wherever platforms are accessed.[20] Similarly, whereas local courts can only order platforms to block material accessed in their jurisdiction, the industry database has the potential to result in worldwide censorship.

---

15 Mark Scott, *Europe Presses American Tech Companies to Tackle Hate Speech*, N.Y. TIMES (Dec. 6, 2016), https://www.nytimes.com/2016/12/06/technology/europe-hate-speech-facebook-google-twitter.html; Amar Toor, *UK Lawmakers Say Facebook, Google, and Twitter Are 'Consciously Failing' to Fight ISIS Online*, VERGE (Aug. 26, 2016), http://www.theverge.com/2016/8/26/12656328/facebook-google-twitter-isis-propaganda-uk-report.

16 Article 19 of the International Covenant on Civil and Political Rights allows states to limit freedom of expression under circumstances that satisfy proportionality review. International Covenant on Civil and Political Rights art. 19, Dec. 19, 1966, 999 U.N.T.S. 172.

17 Code of Conduct on Countering Illegal Hate Speech Online, http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf.

18 Casey Newton, *Facebook, Microsoft, Twitter, and YouTube Are Creating a Database of "Terrorist Content,"* VERGE (Dec. 5, 2016), https://www.theverge.com/2016/12/5/13849570/facebook-microsoft-twitter-google-terrorist-content-database.

19 Unlike in the EU, in the United States, threatening to regulate protected speech implicates the protections of the First Amendment. *See* Fairley v. Andrews, 578 F.3d 518, 525 (7th Cir. 2009) ("Threatening penalties for future speech goes by the name 'prior restraint,' and a prior restraint is the quintessential first-amendment violation."). For instance, the Sheriff of Cook County, Illinois, wrote letters to credit card companies demanding that they prohibit advertisers from using cards to purchase advertisements on Backpage.com since ads might be used for illegal sex-related products or services. Backpage responded by seeking a preliminary injunction against the Sheriff for violating its First Amendment rights. *See* Backpage.com, LLC v. Dart, 807 F.3d 229, 230 (7th Cir. 2015). The court held that the Sheriff had irreparably harmed the site by threatening coercive state action against credit card companies for facilitating speech on the site. *Id.* at 239. The court directed the trial court to issue a temporary injunction ordering the Sheriff to "take no actions, formal or informal, to coerce or threaten credit card companies, processors, financial institutions, or other third parties with sanctions intended to ban credit card or other financial services from being provided to Backpage.com." *Id.*

20 Klonick, *supra* note 6, at 51.

All of this might enjoy some justification if EU regulators focused their efforts on speech proscribed in their countries.[21] But this has not been the case. Calls to remove hate speech have quickly ballooned to cover expression that does not violate existing European law, including "online radicalization" and "fake news."[22] EU officials have pressed a view of hate speech that can be extended to political dissent and newsworthy developments. At risk is censorship creep on a global scale.

Scholarship has explored how formal legal requirements and informal government pressure can result in collateral censorship, the silencing of private actors by other private actors.[23] Free speech scholar and Yale Information Society Project founder Jack Balkin recently warned:

> Currently the Internet is mostly governed by the values of the least censorious regime—that of the United States. If nation states can enforce global filtering, blocking, and delinking, the Internet will eventually be governed

21 I say "some justification" deliberately. European countries have a far different approach to free expression than the United States. *See* FLOYD ABRAMS, THE SOUL OF THE FIRST AMENDMENT xvii, 39, 41–43 (2017). As James Weinstein put it to me, the EU does not really even have a concept of protected speech.

22 Cara McGoogan, *EU Accuses Facebook and Twitter of Failing to Remove Hate Speech*, TELEGRAPH (Dec. 5, 2016), http://www.telegraph.co.uk/technology/2016/12/05/eu-accuses-facebook-twitter-failing-remove-hate-speech/. In late 2017, Germany passed a law levying steep fines for companies' failure to remove hateful conduct and "fake news" within twenty-four hours. *See infra* note 88 (discussing the law). It is, however, unclear if the statute applies to "fake news" outside the realm of hateful ideas targeted at protected groups. *Id.*

23 Michael Meyerson coined the phrase "collateral censorship." Michael I. Meyerson, *Authors, Editors, and Uncommon Carriers: Identifying the "Speaker" Within the New Media*, 71 NOTRE DAME L. REV. 79, 118 (1995). Collateral censorship involves state action because the "government has created incentives for private parties to censor each other." J.M. Balkin, Essay, *Free Speech and Hostile Environments*, 99 COLUM. L. REV. 2295, 2299 (1999). For scholarship exploring collateral censorship resulting from both formal and informal government action, see, for example, REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM xxii (2012); Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2298, 2303–11 (2014) (exploring the difference between old-school speech regulations involving ex ante laws and new-school techniques involving ex post efforts short of threats of legal intervention, such as urging companies to stop doing business with private actors); Derek E. Bambauer, *Orwell's Armchair*, 79 U. CHI. L. REV. 863 (2012) (arguing that hard censorship is preferable to opaque and unaccountable efforts at censorship, such as government's use of unrelated laws as a pretext to block material, payment for filtered access, and ad hoc efforts to persuade intermediaries to restrict content); Seth F Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. PA. L. REV. 11 (2006) (contending that enforcement of copyright, material support, and other laws against intermediaries facilitates censorship by proxy); Christina Mulligan, *Technological Intermediaries and Freedom of the Press*, 66 SMU L. REV. 157 (2013) (arguing that collateral censorship threatens the freedom of the press); Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293 (2011).

> by the most censorious regime. This will undermine the global public good of a free Internet.[24]

The assault on the "global public good of a free Internet" is already underway. As this Article shows, digital expression is conforming to EU speech norms with extremist and hateful speech as catalysts.

This Article has three parts. Part I exposes the pressure facing technology companies to tailor their speech policies to EU norms. As Part I shows, Silicon Valley's recent retreat from a strong commitment to free speech has more to do with compulsion than choice. Part II explores the fallout, highlighting the risk of censorship creep on a global scale. Part III offers safeguards designed to contain extralegal pressure for the good of free expression.

I. THE EU'S POWER OVER PRIVATE SPEECH RULES

After a spate of deadly terror attacks and hate crimes in 2015, European lawmakers told social media companies that they were partly to blame for the violence.[25] In their view, online platforms had enabled violent extremists by giving them access to potential recruits.[26] European lawmakers warned companies that they would face onerous criminal and civil penalties unless online extremism was eliminated.[27]

After the Charlie Hebdo attack, French President François Hollande called for legislation that would make social media platforms criminally liable for users' "extremist" content.[28] French Interior Minister Bernard Cazeneuve followed that warning with meetings in Silicon Valley.[29] Discussions with tech executives bore some fruit: several companies agreed to continue removing terror-related content.[30] As this Part explores, this was just the start of Silicon Valley's concessions to European regulators.

---

24 Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV. (forthcoming 2018) (manuscript at 60–61).

25 *See* Lizzie Plaugic, *France Wants to Make Google and Facebook Accountable for Hate Speech*, VERGE (Jan. 27, 2015), https://www.theverge.com/2015/1/27/7921463/google-facebook-accountable-for-hate-speech-france.

26 *See* Clark, *supra* note 14.

27 Plaugic, *supra* note 25.

28 *Id.*

29 Scott Higham & Ellen Nakashima, *Why the Islamic State Leaves Tech Companies Torn Between Free Speech and Security*, WASH. POST (July 16, 2015), https://www.washingtonpost.com/world/national-security/islamic-states-embrace-of-social-media-puts-tech-companies-in-a-bind/2015/07/15/0e5624c4-169c-11e5-89f3-61410da94eb1_story.html.

30 *Id.*

A. Code of Conduct

On December 3, 2015, the European Commission[31] established the European Internet Forum (the "Forum").[32] The goal was the development of "a joint, voluntary approach" for the detection and removal of "online terrorist incitement and hate speech."[33] Participants included European officials, Europol, and tech companies Facebook, Microsoft, Twitter, and YouTube ("Tech Companies"). The European Commissioner for Migration, Home Affairs, and Citizenship remarked:

> Terrorists are abusing the internet to spread their poisonous propaganda: that needs to stop. The voluntary partnership we launch today with the internet industry [aims] to address this problem. We want swift results. This is a new way to tackle this extremist abuse of the internet, and it will provide the platform for expert knowledge to be shared, [and] for quick and operational conclusions to be developed . . . .[34]

The Forum produced results in short order. On May 31, 2016, the European Commission announced an agreement with the Tech Companies entitled "Code of Conduct on Countering Illegal Hate Speech Online" ("hate-speech agreement" or "the Code").[35] The Tech Companies agreed to prohibit "hateful conduct," defined as speech inciting violence or hatred against protected groups.[36] Reports of hate speech would be reviewed within twenty-

---

31 The European Commission, which represents the common EU interest, is the EU's executive arm. Anu Bradford, *The Brussels Effect*, 107 NW. U. L. REV. 1, 13 (2012). It "ensures that the regulations and directives adopted by the Council [of the European Union] and the Parliament are implemented in the member states." *Id.*

32 EUROPEAN COMM'N, A EUROPEAN AGENDA ON SECURITY: STATE OF PLAY: DECEMBER 2016 (Dec. 21, 2016), https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-security/fact-sheets/docs/20161221/european_agenda_on_security_state_of_play_21122016_en.pdf.

33 European Commission Press Release IP/15/6243, EU Internet Forum: Bringing Together Governments, Europol and Technology Companies to Counter Terrorist Content and Hate Speech Online (Dec. 3, 2015), http://europa.eu/rapid/press-release_IP-15-6243_en.htm.

34 *Id.*

35 Clark, *supra* note 14. Although civil society organizations participated in early meetings held by the European Internet Forum, they were excluded from the negotiations that resulted in the Code. *EDRi and Access Now Withdraw from the EU Commission IT Forum Discussions*, EDRI (May 31, 2016), https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions. As the civil society group European Digital Rights ("EDRi") explained, the European Commission refused to give the groups access to the negotiations and drafts of the agreement. Maryant Fernández Pérez, *New Documents Reveal the Truth Behind the Hate Speech Code*, EDRI (Sept. 7, 2016), https://edri.org/new-documents-reveal-truth-behind-hate-speech-code.

36 European Commission Press Release IP/16/1937, European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech (May 31, 2016), http://europa.eu/rapid/press-release_IP-16-1937_en.htm. That definition was drawn from the 2008 European Framework Decision 2008/913/JHA. *Id.*

four hours and removed if the speech violated companies' terms of service.[37] The European Commission made clear that it would conduct periodic reviews of the Tech Companies' compliance with the hate-speech agreement.[38] The European Commissioner for Justice, Consumers and Gender Equality Vera Jourová hailed the hate-speech agreement as essential to combating the use of social media to "radicalize young people and to spread violence and hatred."[39]

In December 2016, the European Commission issued its first assessment of the Tech Companies' handling of hate-speech reports, and the feedback was not positive. Over a six-week period, twelve organizations, working on behalf of the Commission, reported alleged incidents of hate speech and tracked the companies' response.[40] The European Commission criticized the Tech Companies' "success rate"—the number of requests resulting in removal—and timeliness.[41] Only forty percent of hate-speech reports were reviewed in twenty-four hours, and twenty-eight percent of speech reported as hateful conduct was removed.[42]

In the estimation of the European Commission, the Tech Companies had fallen short on their commitments. Jourová warned that if the Tech Companies "want to convince me and the ministers that the non-legislative approach can work, they will have to act quickly and make a strong effort in the coming months."[43] In other words, more hateful speech needed to be removed, and faster, or else.[44]

B. Blacklist Database

EU authorities have been in contact with social media companies about terrorist groups' use of their services for the past seven years.[45] For years, the contact proceeded on an ad hoc basis with law enforcement asking compa-

37 Under the Code, the Tech Companies would review reports of alleged hate speech on the basis of their terms of service and only where necessary on the basis of law. It is unclear if the Tech Companies would look to the law to determine whether hate speech should be removed. Fernández Pérez, *supra* note 35.

38 *Id.*

39 Amar Toor, *Facebook, Twitter, Google, and Microsoft Agree to EU Hate Speech Rules*, VERGE (May 31, 2016), http://www.theverge.com/2016/5/31/11817540/facebook-twitter-google-microsoft-hate-speech-europe.

40 EUROPEAN COMM'N, CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE: FIRST RESULTS ON IMPLEMENTATION 1 (Dec. 2016), http://ec.europa.eu/information_society/newsroom/image/document/2016-50/factsheet-code-conduct-8_40573.pdf.

41 *Id.* at 3–4.

42 *Id.* at 4.

43 Clark, *supra* note 14.

44 *Id.*

45 Jason Murdock, *Isis: UK Cyber Cops Ramp up Campaign to Curb the Spread of Daesh-Inspired Propaganda*, INT'L BUS. TIMES (Apr. 18, 2016), http://www.ibtimes.co.uk/isis-uk-cyber-cops-ramp-campaign-curb-spread-daesh-inspired-propaganda-1555463.

nies to take down content.[46] The United Kingdom established a Counter Terrorism Internet Referral Unit (CTIRU) to identify and report "violent and nonviolent extremism" to online platforms.[47] From 2010 to 2015, CTIRU secured the removal of 249,091 pieces of terrorist-related content.[48] According to UK officials, CTIRU had no need to file formal notice and takedown requests because tech companies were so cooperative.[49]

Given the success of the CTIRU's efforts, Europol established its Internet Referral Unit, described as a "partnership[ ] with the private sector (to promote 'self-regulation' by online service providers)."[50] Ninety-one percent of the content reported has been removed.[51] Europol has described the private sector's compliance with their removal requests as "voluntary."[52]

In late 2015, social media companies faced growing pressure to systematize the removal process.[53] Extremist content needed to be removed faster and, if possible, before ever appearing online. One suggestion was the adoption of an industry database enabling the detection of banned violent terrorist images, audio, and video files.[54] The database would collect hashes—or

46 Susan Benesch & Rebecca MacKinnon, *The Innocence of YouTube*, FOREIGN POL'Y (Oct. 5, 2012), http://foreignpolicy.com/2012/10/05/the-innocence-of-youtube/.

47 Scott Craig & Emma Llansó, *Pressuring Platforms to Censor Content Is Wrong Approach to Combatting Terrorism*, CTR. FOR DEMOCRACY & TECH. (Nov. 5, 2015), https://cdt.org/blog/pressuring-platforms-to-censor-content-is-wrong-approach-to-combatting-terrorism/.

48 *CTIRU Statistics at a Glance, 250,000th Piece of Online Extremist/Terrorist Material to Be Removed*, METRO. POLICE (Dec. 23, 2016), http://news.met.police.uk/news/250000th-piece-of-online-extremist-slash-terrorist-material-to-be-removed-208698; *see also* Murdock, *supra* note 45. YouTube has granted UK security officials special powers as "super flaggers" of content that violates its terms of service to ensure that extremist content is instantly removed. Sam Jones, *UK Security Services Face Rise of Extremist Content Online*, FIN. TIMES (Sept. 30, 2014), https://www.ft.com/content/a0266d5e-489e-11e4-9d04-00144feab7de.

49 *See* 747 Parl Deb HL (5th ser.) (2013) col. 47 (UK). Police can demand the removal of content that incites or glorifies terrorist acts under the UK's Terrorism Act of 2006. Terrorism Act 2006, c. 11, § 21 (UK). If companies fail to remove terrorist content, they can be charged with encouraging terrorism. *Id.* § 1.

50 EUROPOL, EU INTERNET REFERRAL UNIT: YEAR ONE REPORT 3, https://www.europol.europa.eu/sites/default/files/documents/eu_iru_1_year_report_highlights.pdf.

51 Jennifer Baker, *Europol's Online Censorship Unit Is Haphazard and Unaccountable Says NGO*, ARS TECHNICA (July 4, 2016), https://arstechnica.com/tech-policy/2016/07/europol-iru-extremist-content-censorship-policing/. Although Europol established the Internet Referral Unit in the spring of 2015, it was officially sanctioned via regulation in May 2016. *Id.*; *see also* Press Release, Europol, Europol's Internet Referral Unit to Combat Terrorist and Violent Extremist Propaganda (July 1, 2015), https://www.europol.europa.eu/newsroom/news/europol%E2%80%99s-internet-referral-unit-to-combat-terrorist-and-violent-extremist-propaganda.

52 Baker, *supra* note 51.

53 *Diane Rehm: Growing Pressure on Social Media Sites to Monitor and Remove Terrorist Content*, AM. UNIV. RADIO (Dec. 8, 2015), https://dianerehm.org/shows/2015-12-08/growing-pressure-on-social-media-sites-to-monitor-and-remove-terrorist-content.

54 Kaveh Waddell, *A Tool to Delete Beheading Videos Before They Even Appear Online*, ATLANTIC (June 22, 2016), https://www.theatlantic.com/technology/archive/2016/06/a-tool-to-

unique digital fingerprints—of banned content so it could be instantly flagged and removed.[55]

Silicon Valley initially rejected the idea.[56] Executives expressed concern that valuable content could find its way into the database since there wasn't clear consensus about the meaning of "terrorist speech."[57] Another worry was that governments might try to use the database to silence critics.[58]

Six months later, the Tech Companies reversed course and announced plans for an industry database to help prevent the spread of "violent terrorist imagery."[59] In a press release, the Tech Companies noted that other platforms would be welcome to participate as soon as the database was up and running.[60] The European Commission hailed the database as the next logical step in a public-private partnership to combat extremism.[61] In its view, the database was "a significant step forward in the form of a collaborative

---

delete-beheading-videos-before-they-even-appear-online/488105/. The suggestion drew inspiration from the Microsoft system used to detect and remove child pornography. *Id.*

55 Ellen Nakashima, *There's a New Tool to Take Down Terrorism Images Online. But Social-Media Companies Are Wary of It*, WASH. POST (June 21, 2016), https://www.washingtonpost.com/world/national-security/new-tool-to-take-down-terrorism-images-online-spurs-debate-on-what-constitutes-extremist-content/2016/06/20/0ca4f73a-3492-11e6-8758-d58e76e11b12_story.html?utm_term=.28394b560bb2; *see also* Jamie Condliffe, *Facebook and Google May Be Fighting Terrorist Videos with Algorithms*, MIT TECH. REV. (June 27, 2016), https://www.technologyreview.com/s/601778/facebook-and-google-may-be-fighting-terrorist-videos-with-algorithms/ ("[Hashing] is a mathematical operation that takes a long stream of data of arbitrary length, like a video clip or string of DNA, and assigns it a specific value of a fixed length, known as a hash. The same files or DNA strings will be given the same hash, allowing computers to quickly and easily spot duplicates.").

56 Nakashima, *supra* note 55.

57 *Id.* There was, however, some support for the effort when the idea was initially discussed with tech companies. Microsoft, for instance, announced that it was "providing funding and technical support to Dartmouth College computer scientist Hany Farid," who worked with Microsoft to develop PhotoDNA (which hashes images of child pornography), to "develop a technology to help stakeholders identify copies of patently terrorist content." Reuters, *Google, Facebook Quietly Move Toward Automatic Blocking of Extremist Videos*, FORTUNE (June 26, 2016), http://fortune.com/2016/06/26/google-fb-block-extremist-videos/.

58 Nakashima, *supra* note 55.

59 *Partnering to Help Curb Spread of Online Terrorist Content*, FACEBOOK NEWSROOM (Dec. 5, 2016), http://newsroom.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/; *see also* Sarah Perez, *Facebook, Microsoft, Twitter and YouTube Collaborate to Remove "Terrorist Content" from Their Services*, TECHCRUNCH (Dec. 5, 2016), https://techcrunch.com/2016/12/05/facebook-microsoft-twitter-and-youtube-collaborate-to-remove-terrorist-content-from-their-services/. The Tech Companies made this announcement the day before the European Commission released its critical report demanding stronger action against hate speech. Scott, *supra* note 15.

60 Glyn Moody, *Internet Giants Will Join Forces to Stop Online Sharing of Terrorist Material*, ARS TECHNICA (Dec. 6, 2016), https://arstechnica.com/tech-policy/2016/12/twitter-facebook-microsoft-youtube-terrorist-material-removal/.

61 European Commission Press Release IP/16/4328, EU Internet Forum: A Major Step Forward in Curbing Terrorist Content on the Internet (Dec. 8, 2016), http://europa.eu/rapid/press-release_IP-16-4328_en.htm.

industry response in protecting their users from terrorist content."[62] Google similarly noted its hope that the collaboration would "lead to greater efficiency as we continue to enforce our policies to help curb the pressing global issue of terrorist content online."[63]

The Tech Companies, however, did issue some guidelines.[64] The database would share hashes of "the most extreme and egregious terrorist images and videos . . . content most likely to violate all of our respective companies' content policies."[65] Hashed material would not be immediately deleted from participants' sites.[66] Instead, each company would review content included in the database under its own policies.[67] The Tech Companies acknowledged the importance of clear guidelines to ensure that the industry database would not be used to censor public discourse.[68]

C. More Coercion than Choice

European laws have influenced policy all over the globe, often in ways that strengthen legal regulation.[69] Most obviously, the EU supplies global standards when foreign regulations mimic its strict rules. The EU's influence is also felt when companies adopt its strict standards as a result of market forces.[70] In what has been called the "Brussels Effect," companies follow EU laws in their global operations because it allows them to operate simultaneously in the world's largest internal market and in jurisdictions with more lax standards.[71] In areas like data privacy and antitrust, companies have con-

62 *Id.*

63 Scott Shackford, *How Long Before This Tool to Censor Images from Terrorists Gets Misused?*, REASON (Dec. 6, 2016), http://reason.com/blog/2016/12/06/how-long-before-this-tool-to-censor-imag.

64 *Partnering to Help Curb Spread of Online Terrorist Content*, *supra* note 59.

65 Liat Clark, *Facebook, Twitter, Microsoft, YouTube Launch Shared Terrorist Media Database*, WIRED UK (Dec. 6, 2016), http://www.wired.co.uk/article/facebook-twitter-microsoft-youtube-launch-shared-terrorism-database.

66 *Id.*

67 *Id.* It is unclear from the announcement whether participants in the database would provide a way for users to challenge inclusion of their content in the hash database. Also, because the database enforces TOS violations, it would likely result in worldwide blocking of hashed material, as is commonly true of other TOS violations. *See, e.g.*, *User Content and Conduct Policy*, GOOGLE, https://www.google.com/intl/en_us/+/policy/content.html (last visited Nov. 14, 2017).

68 Moody, *supra* note 60.

69 *See* Bradford, *supra* note 31 (exploring the dynamics of Vogel's California Effect in a global context). David Vogel documented the role that California laws had on other states in much the same manner. *See* DAVID VOGEL, TRADING UP: CONSUMER AND ENVIRONMENTAL REGULATION IN A GLOBAL ECONOMY 259–60 (1995). I explored the California Effect in my work on the privacy policymaking of state attorneys general. *See* Danielle Keats Citron, *The Privacy Policymaking of State Attorneys General*, 92 NOTRE DAME L. REV. 747 (2016).

70 Bradford, *supra* note 31, at 6.

71 *Id.* (explaining that the Brussels Effect is most likely to occur when EU regulation governs policy areas of low elasticity).

formed their global operations to more restrictive European standards to avoid the expense of piecemeal compliance.[72]

Market forces, however, do not explain Silicon Valley's recent adoption of EU speech norms related to extremist speech. The Tech Companies did not change their speech policies and practices for efficiency's sake. They were not trying to attract advertising fees or advocates' approval. They were responding to government threats, not voluntarily making changes to their policies.

Tech companies yielded to European regulators' demands because they knew their threats were not idle. In the European Union, unlike in the United States, there isn't a heavy presumption against speech restrictions.[73] Hate speech and extremist expression were already banned in some EU member states.[74] Social media providers knew that if onerous laws were

72 *Id.* at 18. As Paul Schwartz insightfully explains, the trend is towards the harmonization of EU-U.S. privacy policies, most recently with the Privacy Shield. Paul M. Schwartz, *The EU-U.S. Privacy Collision: A Turn to Institutions and Procedures*, 126 HARV. L. REV. 1966 (2013). Continued harmonization is not assured if the Trump administration calls into question the deal struck in the Privacy Shield. *See* Paul M. Schwartz & Karl-Nikolaus Peifer, *Transatlantic Data Privacy Law*, 106 GEO. L.J. 115 (2017). If that is the case, we shall see an upward regulatory trajectory along the lines of the Brussels Effect: global companies will likely endeavor to comply with the EU's strict General Data Protection Rule, which goes into effect in May 2018.

73 In the EU, laws penalizing speech must satisfy a proportionality analysis. *Report from the Commission on Subsidiarity and Proportionality*, (Mar. 15, 2010), http://ec.europa.eu/dgs/secretariat_general/relations/relations_other/npo/subsidiarity_reports_en.htm. Under Article 20 of the Covenant on Political and Civil Rights, "incitement to discrimination, hostility or violence" on the basis of "national, racial or religious hatred" is prohibited. G.A. Res. 2200A (XXI), at 20 (Dec. 16, 1966).

74 A number of EU member states outlaw hate speech and Holocaust denial. Leonid Bershidsky, Opinion, *Europe Gets U.S. Tech Leaders to Self-Censor*, BLOOMBERG VIEW (May 31, 2016), https://www.bloomberg.com/view/articles/2016-05-31/europe-gets-u-s-tech-leaders-to-self-censor. In 2009, the European Court of Human Rights upheld the conviction of a member of the Belgian Parliament for inciting racial discrimination by distributing leaflets calling for a "Belgians and Europeans *First*" policy and saying, "Stand Up against the Islamification of Belgium." Abrams, *supra* note 21, at 42.

Similar efforts to regulate hate speech in the United States would surely fail given strict scrutiny review of viewpoint-based speech regulations. *Id.* at xvii, 39–40 (exploring American approach to free speech in which statements, whatever their offense to groups, are routinely protected from viewpoint-based regulation, such as hate speech bans); *see also* Snyder v. Phelps, 562 U.S. 443, 455, 458 (2011) (vacating judgment for intentional infliction of emotional distress resulting from protestors' bigoted messages because such commentary amounted to "speech on a matter of public concern" and could not be penalized "simply because it is upsetting or arouses contempt" (quoting Connick v. Myers, 461 U.S. 138, 146 (1983))). That is why U.S. officials have proceeded with a light touch in their efforts to influence Silicon Valley's decisions about extremist speech. After the pipe bomb explosions in New York and New Jersey in 2016, President Obama urged tech companies to help "to push back against online extremist content and all messages of hate." Press Release, The White House, Office of the Press Secretary, Statement by the President on the Explosions in New York City and New Jersey (Sept. 19, 2016). That message came with an explicit recognition that the White House was not considering legal intervention. As the

passed and enforced, hosting user-generated content would be prohibitively expensive. No matter how often EU lawmakers describe the recent changes to private speech practices as "voluntary" decisions, they can only be reasonably understood as the product of government coercion.

Comparing the dynamics of earlier changes in companies' speech policies to recent developments demonstrates the point. In interviews conducted in 2010 and 2011, safety personnel explained that their platforms banned hate speech because doing so comported with their employers' sense of corporate social responsibility and the wishes of advertisers and advocacy groups.[75] Facebook, Microsoft's Xbox Live, and other content hosts adopted different definitions of hate speech,[76] from the narrow (e.g., speech targeting marginalized groups with violence) to the broad (e.g., speech that demeans marginalized groups).[77] Corporate hate speech policies were a reflection of business choices, not government pressure.[78]

Similarly, changes in the private speech rules around nonconsensual pornography were a reflection of market forces. As early as 2011, advocacy groups were pressing platforms to change their rules regarding the posting of nude images without subjects' consent.[79] They argued that nonconsensual pornography was neither good for business nor the result of voluntary sexual expression.[80] Advocates' only leverage was their ability to rally public opinion, which could impose modest costs if some users and advertisers

White House Press Secretary recognized, "There are obviously a lot of complicated First Amendment issues and other things." Press Release, The White House, Office of the Press Secretary, Press Briefing by Press Secretary Josh Earnest (Jan. 8, 2016).

75 Citron & Norton, *supra* note 7, at 1454. Beginning in 2010, the Anti-Cyber Hate Working Group (of which I was a member) brought together tech companies, advocates, and academics to discuss best practices regarding hate speech. *See* Jeffrey Rosen, *The Delete Squad*, NEW REPUBLIC (Apr. 29, 2013), https://newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules (describing the Anti-Cyber Hate Working Group's meetings and efforts).

76 Alexandra Burlacu, *Microsoft Tackles Online Hate Speech with New Tools and Resources to Combat Abuse*, TECH TIMES (Aug. 27, 2016), http://www.techtimes.com/articles/175274/20160827/microsoft-tackles-online-hate-speech-with-new-tools-and-resources-to-combat-abuse.htm; Brian Heater, *What Facebook, Twitter, Youtube and Others Are Doing to Tackle Hate Speech*, TECHCRUNCH (Aug. 15, 2017), https://techcrunch.com/2017/08/15/what-facebook-twitter-youtube-and-others-are-doing-to-tackle-hate-speech/.

77 Citron & Norton, *supra* note 7, at 1468 n.185. Until 2016, Twitter refused to ban hate speech because it feared censoring such a broad category of expression. Kate Benner, *Twitter Adds New Ways to Curb Abuse and Hate Speech*, N.Y. TIMES (Nov. 15, 2016), https://www.nytimes.com/2016/11/16/technology/twitter-adds-new-ways-to-curb-abuse-and-hate-speech.html?_r=0.

78 Citron & Norton, *supra* note 7, at 1453–54.

79 Danielle Keats Citron, *Online Engagement on Equal Terms*, B.U. L. REV. ANNEX (Oct. 19, 2015), https://www.bu.edu/bulawreview/citron-online-engagement-on-equal-terms/ (discussing Twitter and other platforms' changes to speech rules to ban nonconsensual pornography).

80 *See generally* CITRON, *supra* note 6; *see also* Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345 (2014); Franks, *supra* note 11, at 10–11.

exited their platforms. After prominent individuals decried the nonconsensual posting of their nude images and popular media outlets criticized companies' inattention to the problem in 2014, the major online platforms overhauled their policies to ban the practice.[81] To be sure, advocacy groups had the ability to rally public support behind the notion that companies should ban nonconsensual pornography. But, unlike governments, they had neither the leverage nor the power to make it prohibitively expensive for companies to ignore their advocacy.[82] That is why it took considerable time—and crucially, market pressure—for platforms to change their positions on nonconsensual pornography.

By contrast, when the Tech Companies changed their speech policies and procedures regarding hateful and extremist material in 2016, they did so in the shadow of threatened regulation.[83] In what was surely timed to lessen the blow of the European Commission's critical review of their compliance with the hate-speech agreement, the Tech Companies released their announcement about the industry database the day before the European Commission released its report.[84] As Eve Peyser has observed, the industry database was surely the "quick and strong effort the EU asked for."[85] EU regulators wielded their power to impose material costs on extremist speech to pressure conformity with their speech norms.

The demands of European leaders have only escalated since the announcement of the hate-speech agreement and the industry database. After a series of terrorist attacks in London in 2017, British Prime Minister Theresa May and French President Emmanuel Macron threatened to fine tech companies for "fail[ing] to remove 'extremist' propaganda from their platforms."[86] They called upon social media platforms to use automation to prevent publication, rather than relying on users to flag content for deletion.[87] Germany made good on its threats by passing a law sanctioning fines

81 Franks, *supra* note 11, at 3–6.

82 As discussed above, advocacy groups could not support lawsuits against online platforms for user-generated content given their immunity from liability under Section 230 of the Communications Decency Act. *See supra* note 7 and accompanying text.

83 Gretel Kauffman, *EU Urges Social Media Giants to Act on Hate Speech*, CHRISTIAN SCI. MONITOR (Dec. 5, 2016), http://www.csmonitor.com/Technology/2016/1205/EU-urges-social-media-giants-to-act-on-hate-speech.

84 Scott, *supra* note 15. The Tech Companies surely had advance knowledge of the impending report—the press release about the shared database was perhaps meant to blunt the Commission's criticism.

85 Eve Peyser, *Twitter and Facebook Randomly Crack Down on Terrorist Videos After EU Warning*, GIZMODO (Dec. 5, 2016), http://gizmodo.com/twitter-and-facebook-decide-to-crack-down-on-spread-of-1789710389.

86 Amanda Paulson & Eva Botkin-Kowacki, *In Terror Fight, Tech Companies Caught Between US and European Ideals*, CHRISTIAN SCI. MONITOR (June 23, 2017), https://www.csmonitor.com/Technology/2017/0623/In-terror-fight-tech-companies-caught-between-US-and-European-ideals.

87 Cabinet Office and Home Office, French-British Action Plan: Internet Security, 2017 (UK), https://www.gov.uk/government/publications/french-british-action-plan-internet-security.

up to fifty-seven million euros if companies fail to remove hate speech within twenty-four hours of its reporting.[88]

In response, Google announced a four-part plan to address terrorist propaganda including increased use of technology to identify terrorist-related videos, hiring additional content moderators, removing advertising on potentially objectionable videos, and directing potential terrorist recruits toward counter-radicalization videos.[89] Facebook announced its use of artificial intelligence to stop the spread of terrorist propaganda and hiring of 3000 more people to review speech reported as terms-of-service (TOS) violations.[90] These efforts appear designed to prevent European leaders from adopting or enforcing legislation that would punish tech companies for failing to censor extremist content adequately.[91]

II. Censorship Creep

Without question, EU pressure to remove hateful conduct and violent terrorist content could be beneficial. Changes in companies' speech practices could prevent disenfranchised individuals from seeing Al-Qaeda videos calling for the death of Jews and then going to synagogues with guns. The shared database could facilitate the swift removal of gruesome beheading videos, preventing their viral spread all over the internet. It could ensure that a "kill list" of U.S. soldiers is removed before anyone has a chance to see it.[92] With less terrorist propaganda and less hate speech online,[93] there

88 Melissa Eddy & Mark Scott, *Delete Hate Speech or Pay up, Germany Tells Social Media Companies*, N.Y. Times (June 30, 2017), https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html; Greenberg Traurig LLP, *German Government Passes Disputed Draft of Act Improving Law Enforcement on Social Networks*, Lexology (Apr. 28, 2017), http://www.lexology.com/library/detail.aspx?g=D6d06c09-9f8b-43ed-8f2d-761c5199befc; Oliver Süme, *Germany: Draft "Network Enforcement Law" to Tackle Hate Speech and Fake News*, Fieldfisher (Apr. 5, 2017), http://www.fieldfisher.com/publications/2017/04/germany-draft-network-enforcement-law-to-tackle-hate-speech-and-fake-news. The German law also apparently covers "fake news." *Id.* As Paula Kift explains, the German law "does not define 'fake news.' Instead, it suggests that 'fake news' [is] indictable only when [it] 'objectively' meet[s] any of the elements of the offenses." E-mail from Paula Kift to Danielle Citron (June 3, 2017, 1:43 PM) (on file with author) (translating German law into English and explaining its meaning).

89 Kent Walker, *Four Steps We're Taking Today to Fight Terrorism Online*, Google (June 18, 2017), https://blog.google/topics/google-europe/four-steps-were-taking-today-fight-online-terror/.

90 Monika Bickert & Brian Fishman, *Hard Questions: How We Counter Terrorism*, Facebook Newsroom (June 15, 2017), https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/.

91 *See* Clark, *supra* note 14; Moody, *supra* note 60.

92 This example comes from British ISIS operative Sally Jones who used social media to incite Westerners to violence. She issued a "kill list" of 100 servicemen and U.S. veterans via Twitter. *Radicalization: Social Media and the Rise of Terrorism: Hearing Before the Subcomm. on Nat'l Sec. of the H. Comm. on Oversight and Gov't Reform*, 114th Cong. 6 (2015) (testimony of Mark D. Wallace, CEO of the Counter Extremism Project) [hereinafter *Radicalization Social Media*].

might be fewer people joining ISIS fighters in Syria or planting bombs in shopping markets or houses of worship.[94]

Although these potential benefits are not in doubt, there are potential costs as well. Companies' TOS policies could be interpreted to prohibit speech far beyond speech inciting hatred of (or calling for violence against) vulnerable groups or violent extremist content. They could result in the global deletion of a government official's tweets.[95] They could lead to the worldwide removal of websites criticizing political candidates.[96] They could result in the global suspension of civil rights activists' Facebook profiles.[97]

This Part explains how this could happen. It begins by exploring the concept of censorship creep. It exposes the costs of censorship creep, including the suppression of legitimate debate and counter speech that might convince people to reject bigotry and terrorist ideology.

A. Concept

The term creep refers to "the idea that a tool designed for one purpose ends up being used for another one."[98] Tools or programs designed to accomplish a particular end or to solve a specific problem are gradually extended to other uses or contexts.[99] In the engineering context, the phe-

93 By which I mean hate speech that would likely be adjudicated as illegal under the laws of EU member states.

94 Evidence has suggested that ISIS adherents who committed atrocities have used social media to spread terrorist propaganda and justify violence. *See Radicalization Social Media*, *supra* note 92, at 5.

95 This is a riff on the hate speech conviction of a Belgian Parliament member who distributed leaflets calling for a "Belgians and European First" policy and saying, "Stop the Sham Immigration Policy, Send non-European sub-seekers home." Abrams, *supra* note 21, at 42.

96 Laurel Wamsley, *Austrian Court Rules Facebook Must Delete Hate Speech*, NPR (May 8, 2017), http://www.npr.org/sections/thetwo-way/2017/05/08/527398995/austrian-court-rules-facebook-must-delete-hate-speech.

97 Facebook temporarily banned a well-known Black Lives Matter activist and journalist, Shaun King, who posted racist messages that he received. Sam Levin, *Facebook Temporarily Blocks Black Lives Matter Activist After He Posts Racist Email*, GUARDIAN, (Sept. 12, 2016), https://www.theguardian.com/technology/2016/sep/12/facebook-blocks-shaun-king-black-lives-matter. As Mr. King noted after Facebook restored his service, "SO many of you have told me that you have had your accounts suspended for FIGHTING BIGOTRY while the bigots often seem to be able to say whatever the hell they want." Shaun King, FACEBOOK (Sept. 9, 2016), https://www.facebook.com/shaunking/photos/a.799605230078397.1073741828.799539910084929/1135506076488309/?type=3&theater. Mr. King's experience is not unique. *See* Tracy Jan & Elizabeth Dwoskin, *A White Man Called Her Kids the N-Word. Facebook Stopped Her from Sharing It*, WASH. POST (July 31, 2017), https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html?utm_term=.97d6e7103703.

98 Brett Frischmann & Evan Selinger, Being Human in the Twenty-First Century (forthcoming 2018) (manuscript at 29) (on file with author).

99 *Id.*

nomenon is called function creep.[100] Mission creep is another term capturing this phenomenon.[101]

Depending on the circumstances, creep can have benefits. New uses of general-purpose technologies can produce economic growth, innovation, and other positive externalities.[102] But creep can be costly as well. It can have pernicious effects that are so gradual that they elude careful review.

Consider the creep involved in certain surveillance programs. As Frank Pasquale and I explored in prior work, fusion centers began as federal-state-local partnerships to combat terrorism.[103] Over time, their mission became unmoored from their antiterror beginnings to cover all crimes, threats, and hazards.[104] While scant security was gained with the expansion of fusion centers' work, it came at a cost, including a loss of individual privacy, chilled speech, and a redirection of resources away from core intelligence missions.[105]

I now turn to the concept of censorship creep, which refers to the expansion of speech policies beyond their original goals.[106] As Paul Bernal explains, "when you build a censorship system for one purpose, you can be pretty certain that it will be used for other purposes."[107] Section II.B explores the troubling potential for censorship creep in private speech programs adopted under EU pressure. It explains the reasons for censorship creep and details its costs to individual freedoms and law enforcement.

B. Causes

Several trends will fuel and exacerbate censorship creep—definitional ambiguity, global enforcement of companies' speech rules, and opacity of private speech practices.[108] This Section explores the implications of these trends.

---

100 *Id.*

101 *See Mission Creep*, OXFORD ENGLISH DICTIONARY, http://www.oed.com/view/Entry/119999?redirectedFrom=Mission©reep#eid36371891 (last visited Oct. 29, 2017).

102 FRISCHMANN & SELINGER, *supra* note 98.

103 *See* Danielle Keats Citron & Frank Pasquale, *Network Accountability for the Domestic Intelligence Apparatus*, 62 HASTINGS L.J. 1441, 1463–64 (2011) (exploring the problem of mission creep in the context of fusion centers).

104 *See id.* at 1463.

105 *See id.* at 1446. The Senate Permanent Subcommittee on Investigations largely agreed with this assessment in a report issued in 2012. PERMANENT SUBCOMM. ON INVESTIGATIONS, FEDERAL SUPPORT FOR AND INVOLVEMENT IN STATE AND LOCAL FUSION CENTERS 1 (Oct. 3, 2012), https://www.hsgac.senate.gov/download/report_federal-support-for-and-involvement-in-state-and-local-fusions-centers.

106 *See* Paulson & Botkin-Kowacki, *supra* note 86.

107 Paul Bernal, *Censorship and Surveillance . . .*, PAUL BERNAL'S BLOG (Sept. 25, 2014), https://paulbernal.wordpress.com/2014/09/25/censorship-and-surveillance/.

108 *See* Eugene Volokh, *The Mechanisms of the Slippery Slope*, 116 HARV. L. REV. 1026, 1112 (2003); *see also* ACCESS NOW, ACCESS NOW POSITION PAPER: A DIGITAL RIGHTS APPROACH TO PROPOSALS FOR PREVENTING OR COUNTERING VIOLENT EXTREMISM ONLINE 15 (Nov. 2016), https://www.accessnow.org/cms/assets/uploads/2016/10/CVE-online-10.27.pdf; Jens-

1. Definitional Ambiguity

Censorship creep happens when speech rules are based on ambiguous terminology.[109] Without clear guidelines and specific examples, vague terms are vulnerable to revision and expansion.[110] Consider the Code's definition of "illegal hate speech": speech inciting violence or hatred against a group or a member of such a group based on race, religion, national, or ethnic origin.[111] Inciting hatred against a group is an ambiguous concept. It could be interpreted to cover speech widely understood as hateful, such as describing members of a religious group as vermin responsible for crime and disease. But it could also be understood as covering speech that many would characterize as newsworthy. Given the term's ambiguity, incitement of hatred could extend to criticism of Catholics for covering up priests' sexual exploitation of children.[112] It could be interpreted as applying to speech challenging Islamic fundamentalism for its homophobia or suppression of women.[113] It could be extended to speech exposing hatred faced by racial minorities.[114]

In the context of the hate-speech agreement, censorship creep is not a theoretical possibility. It is already happening. European officials have conflated "illegal hate speech" with terrorist content, extremist speech, and bogus news stories. In criticizing the Tech Companies' implementation of the agreement, the European Commissioner for Justice, Consumers and Gender Equality pointed to the companies' failure to remove "online radicalisation, illegal hate speech [and] fake news."[115] Legitimate debate could easily fall within that broad characterization of hate speech.[116]

Henrik Jeppesen, *First Report on the EU Hate Speech Code of Conduct Shows Need for Transparency, Judicial Oversight, and Appeals*, CDT BLOG (Dec. 12, 2016), https://cdt.org/blog/first-report-eu-hate-speech-code-of-conduct-shows-need-transparency-judicial-oversight-appeals/.

109 *See* Volokh, *supra* note 108, at 1112.

110 This concern animates the overbreadth doctrine in American constitutional law.

111 European Commission Press Release IP/16/1937, *supra* note 36. The Code specifically refers to the Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law as the legal basis for its definition of illegal hate speech. *Id.* In the review commissioned by the European Commission, the twelve NGOs reporting alleged hate speech had widely varied views of hate speech—some NGOs had a removal rate of approximately sixty percent, while others had a removal rate of five percent. *See Code of Conduct on Countering Illegal Hate Speech Online: First Results on Implementation*, *supra* note 40, at 4; *see also* Jeppesen, *supra* note 108.

112 *Cf.* Robert Post, *Hate Speech*, *in* EXTREME SPEECH AND DEMOCRACY 125, 126 (Ivan Hare & James Weinstein eds., 2009).

113 *Cf. id.*

114 *See* Jan & Dwoskin, *supra* note 97 (discussing cases where Facebook suspended accounts or removed speech of users who were sharing anguish over hate speech targeting them but left up speech and accounts of white supremacists responsible for such hate speech).

115 McGoogan, *supra* note 22.

116 *See* Raman Jit Singh Chima, *Beware: Countering "Violent Extremism" Online Risks Human Rights*, ACCESS NOW (Nov. 2, 2016), https://www.accessnow.org/beware-countering-violent-extremism-online-risks-human-rights/.

That vague definitions of hate speech can be used to suppress legitimate dissent is a long-standing concern. In particular, the definition of hate speech featured in the hate-speech agreement has long been subject to this critique. During the drafting of the International Covenant on Civil and Political Rights in the 1940s, the United States (led by Eleanor Roosevelt) argued that dictators could manipulate the same definition of hate speech to censor dissent.[117]

As Roosevelt argued seventy years ago, governments may be tempted to report political dissent as hate speech in the hopes that companies will remove it. Human rights advocate Jacob Mchangama has wondered if tech companies will be able to resist the "inevitable demands" of states to remove content they determine to be "'hateful' or 'extremist.'"[118] Resistance might be difficult, especially if a significant portion of the companies' market share comes from the countries requesting the removal of speech.[119]

The industry database raises similar concerns. The notion of violent terrorist content is vague and thus subject to expansion. Although the Tech Companies have pledged to include only the "most extreme and egregious terrorist images and videos,"[120] what constitutes extreme and egregious terrorist content is unclear. The Tech Companies currently have different definitions of terrorist content in their terms of service, from "content intended to recruit for terrorist organizations"[121] to "[v]iolent threats (direct or indirect)" to groups of people.[122] Without clear parameters about what content will be included, the industry database is "vulnerable to mission creep."[123]

Governments are likely to capitalize on the lack of clarity in the meaning of terrorist content. The Tech Companies may face pressure from state actors to include graphic and violent content of all kinds, not just terrorist imagery. Along these lines, a UK Security and Immigration Minister has argued that platforms should block terrorist content even if it is not illegal because people do not want to see "unsavoury" material.[124] Government

117 *See* Evelyn Aswad, *The Role of U.S. Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban*, 1 COLUM. HUM. RTS. L. REV. ONLINE 1, 6 (2016) (analyzing how hate speech code of conduct is inconsistent with the UN Guiding Principles on Business and Human Rights and the Global Network Initiative).

118 Kauffman, *supra* note 83.

119 As Margot Kaminski thoughtfully reminded me, the Tech Companies are capable of pushing back against requests to block specific content. For instance, Google refused to take down the Innocence of Muslims video in the United States. But the vague terms make it easier for platforms to cave to state pressure and the consequences can be the blocking or removal of content worldwide. *See* Benesch & Mackinnon, *supra* note 46.

120 Perez, *supra* note 59.

121 *Violent or Graphic Content*, YOUTUBE, https://support.google.com/youtube/answer/2802008?hl=EN&ref_topic=2803176 (last visited Oct. 26, 2017).

122 *The Twitter Rules*, *supra* note 10.

123 Emma Llansó, *Takedown Collaboration by Private Companies Creates Troubling Precedent*, CDT BLOG (Dec. 6, 2016), https://cdt.org/blog/takedown-collaboration-by-private-companies-creates-troubling-precedent/.

124 Liat Clark, *UK Gov Wants 'Unsavoury' Web Content Censored*, WIRED UK (Mar. 15, 2014), http://www.wired.co.uk/article/government-web-censorship.

authorities could suggest inclusion of hashed videos of pornography or political protests.[125] Although companies are ultimately in charge of the database, they might include content at a government's request that otherwise they would not.

Supporters of the industry database have pointed to the success of Microsoft's PhotoDNA, which uses hash technology to remove child pornography from the internet.[126] The PhotoDNA database, run by the National Center for Missing and Exploited Children (NCMEC), has not morphed into a tool for censorship of other types of material.[127] But child pornography is vastly different from terrorist imagery. Child pornography is not an ambiguous concept—its features can be easily defined and identified. The child pornography hash database is managed by NCMEC, an organization with expert knowledge and experience in child pornography.[128] As Andy Sellars has explained, child pornography "is really the only place where media is contraband by its very definition."[129]

By contrast, whether content amounts to violent and egregious terrorist material depends on the overall context, including the message and precise wording.[130] As the ACLU's Lee Rowland explains, "Algorithms are not good

125 *See* GEOFFREY R. STONE, PERILOUS TIMES: FREE SPEECH IN WARTIME 555–56 (2004).

126 Waddell, *supra* note 54.

127 Hany Farid, the computer scientist who worked on Microsoft's PhotoDNA, worked with the Counter Extremism Project to develop a hash system for terrorist imagery. Kate Knibbs, *Extreme Moderation*, RINGER (Jan. 17, 2017), https://theringer.com/curbing-terrorist-social-media-activity-facebook-twitter-google-601ff9684068. Farid offered the software to social media companies and was puzzled when they did not want to use it, even though it would do what the companies were "already doing, faster, cheaper, and more efficiently, with less errors." *Id.*

128 Sarah Jeong, *Terror Scanning Database for Social Media Raises More Questions than Answers*, MOTHERBOARD (Dec. 9, 2016), https://motherboard.vice.com/en_us/article/social-media-terror-scanning-database.

129 *Id.* In April 2017, Facebook announced its adoption of new tools to address the problem of nonconsensual pornography. Alex Hern, *Facebook Launching Tools to Tackle Revenge Porn*, GUARDIAN (Apr. 5, 2017), https://www.theguardian.com/technology/2017/apr/05/facebook-tools-revenge-porn. Facebook will allow users to report intimate photos posted without consent to "specially trained representatives" from the site's community operations team who will review reported images and remove them if they violate Facebook's community standards. *Id.* Facebook will use photo-matching technologies to remove images it has already determined were shared without consent and in violation of its community standards. *Id.* The photo-matching technologies will only be used on platforms owned by Facebook. *See* Emma Grey Ellis, *Facebook's New Plan May Curb Revenge Porn, but Won't Kill It*, WIRED (Apr. 6, 2017), https://www.wired.com/2017/04/facebook-revenge-porn/. As I told *Wired* magazine at the time of the announcement, Facebook should educate their content moderators about the possibility of censorship creep, so that the photo-sharing technologies are only applied to nonconsensual pornography. I serve on a small team of safety experts advising Facebook on how to ensure that its efforts squarely address the problem of nonconsensual pornography and do not reach beyond it. *See* Antigone Davis, *The Facts: Non-Consensual Intimate Image Pilot*, FACEBOOK (Nov. 9, 2017), https://newsroom.fb.com/news/h/non-consensual-intimate-image-pilot-the-facts/.

130 *See* Waddell, *supra* note 54.

at determining context and tone like support or opposition, sarcasm or parody."[131] Unlike child pornography's clear illegality, violent terrorist speech may be precisely that, or, on the other hand, it may be news or advocacy against violent ideologies.[132]

2. Global Deletion

The recent changes to private speech policies and practices risk turning deletion into the global default. How so? There are two interrelated reasons—first, deletion may become the fallback response for removal requests; and second, TOS agreements are often applied globally.

To illustrate the likelihood that removal will become automatic, consider the hate-speech code of conduct. Removal may become the fallback position on hate speech to forestall criticism and new regulation.[133] A presumption of deletion is certainly cheaper than bearing the costs of new laws. For this reason, the Secretary General of the Council of Europe warned that private entities in public-private partnerships will likely exercise "excessive control" over online content to avoid liability for the transmission of illegal content.[134]

A default practice of removal may be reinforced by the hate-speech agreement's requirement that the Tech Companies respond to reports of hate speech within twenty-four hours. Speed inevitably sacrifices thoughtful deliberation. The short deadline provides additional reason for removal to become the default response to reported hate speech.[135]

Companies' presumptive deletion of hate speech is bound to have a global impact because TOS agreements are involved rather than court orders or other forms of legal process. TOS agreements are typically the same across the globe.[136] Thus, decisions to delete or block content as TOS violations mean content will be deleted or blocked everywhere the platform is viewed.[137]

---

131 *Id.*

132 *See id.*

133 *See* Jeppesen, *supra* note 108.

134 *Council of Europe Secretary General Concerned About Internet Censorship: Rules for Blocking and Removal of Illegal Content Must Be Transparent and Proportionate*, COUNCIL OF EUR. (June 1, 2016), https://www.coe.int/en/web/tbilisi/-/council-of-europe-secretary-general-concerned-about-internet-censorship-rules-for-blocking-and-removal-of-illegal-content-must-be-transparent-and-prop.

135 *See* Jillian C. York, *European Commission's Hate Speech Deal with Companies Will Chill Speech*, ELECTRONIC FRONTIER FOUND (June 3, 2016), https://www.eff.org/deeplinks/2016/06/european-commissions-hate-speech-deal-companies-will-chill-speech.

136 YouTube's description of its terms of service is the same for inside the United States as outside it. *Terms of Service*, YOUTUBE, https://www.youtube.com/t/terms (last visited Oct. 26, 2017). The same is true for Twitter: its description of its terms of service is the same for inside the United States as outside it. *The Twitter Rules*, *supra* note 10.

137 Interview with Emma Llansó, Director of Free Expression at the Center for Democracy and Technology, Washington, D.C. (Jan. 15, 2017) (notes on file with author).

Consider the different impact of a court order and a TOS violation. Suppose France successfully prosecuted a poster for inciting violence against Muslims. With a court order in hand, French authorities could ask online platforms to block the illegal hate speech. The content would likely be blocked from access in France and only in France. But TOS violations proceed in a more systematic fashion. Terms of service related to hateful and extremist speech apply to the Tech Companies' global operations. Because the hate-speech agreement is operationalized through terms of service, a presumption of removal would mean worldwide removal.

The industry database also has the potential to blacklist content across the world. Companies may feel pressure to remove hashed images that other participants have designated as extremist expression in violation of terms of service.[138] If that were the case, the industry database could become a "delete-it-all" program.[139] Material included in the database might disappear or never appear anywhere online.[140]

Although the current plan is for each participant to conduct an independent review of flagged content, it may change as the Tech Companies and future participants come to trust each other's assessments.[141] Given the dominant market positions of the current participants and the likelihood that companies will follow each other's lead, terrorist content included in the database would be invisible to most online users. If a hash is treated as an automatic reason to block content across the major platforms, then the content would effectively not exist. Such censorship would be impossible to "route[ ] around."[142]

---

138 *Id.*

139 This is a paraphrase of Emma Llansó's comments about Conversation AI, an automated program designed to stop and delete harassing comments. Andy Greenberg, *Inside Google's Internet Justice League and Its AI-Powered War on Trolls*, WIRED (Sept. 19, 2016), https://www.wired.com/2016/09/inside-googles-internet-justice-league-ai-powered-war-trolls/. My analysis focuses on content hosts, the top layer of the internet. All of the concerns that I raise in this piece take on greater significance the more power over speech that companies possess and the fewer the options available to consumers.

140 For an analogous phenomenon in the copyright field involving algorithmic matching and filtering systems like YouTube's Content ID, see Matthew Sag, *Internet Safe Harbors and the Transformation of Copyright Law*, 93 NOTRE DAME L. REV. 499 (2017).

141 *See Dangerous Precedent, Data Quality—Talking Tech w/ Emma Llansó & Aimee Rogstad Guidera*, CDT PODCAST (Dec. 19, 2016), https://soundcloud.com/cdt-tech-talk/dangerous-precedent-data-quality-talking-tech-w-emma-llanso-aimee-rogstad-guidera.

142 John Perry Barlow, *Censorship 2000*, INTERNET SOC'Y, https://www.isoc.org/oti/articles/1000/barlow.html (last visited Oct. 29, 2017). John Gilmore's statement that "[t]he Internet treats censorship as though it were a malfunction and routes around it" made sense in 1992. *See id.* Twenty years ago, there were no companies like Facebook, Google, and Twitter dominating online attention. Online platforms did not have hundreds of millions (let alone billions) of subscribers as do popular social media sites today. Then, content blocked on one site could reappear on another site, and the number of potential readers would not vary too much. But today being blocked on the dominant platforms would considerably change access to content.

This will likely happen when smaller companies are given access to the industry database. Startups and other capital-constrained companies likely lack the resources to conduct a review every time a hash is detected on their services.[143] They may be inclined to remove content included in the database when it appears on their sites.[144] Hashes would then operate as an automatic ban for smaller platforms.

As the Center for Democracy and Technology's Free Expression Director Emma Llansó warns, the industry database "will become a target for governments and private actors seeking to suppress speech across the web."[145] It gives governments a single point of pressure that, if successful, would enable the blocking of all types of disfavored content across the internet. Scott Shackford noted: "Once a tool can be used to censor, *en masse*, a violent photo from some terrorist of the Islamic State, that tool can be used to censor anything in similar broad strokes."[146]

3. Opacity

Compounding these concerns is the opacity of private speech practices. Hateful or terrorist content is being removed outside of formal governmental processes.[147] When EU authorities file reports of hate speech under terms of service, it is through a company's complaint system rather than an administrative or judicial process.[148] Unlike requests before an administrative or judicial body, there are no official levers to find out that a government has made the request.[149]

In recent years, the dominant online platforms have provided transparency reports that let users know the number of formal requests made by specific countries to take down content.[150] But those transparency reports may not fully capture governmental requests to remove hate speech or to add violent extremist content to the shared industry database. Because companies are being asked to remove hateful or extremist speech through their private processes rather than governmental ones, there is no guarantee of transparency. As Part III addresses, those requests may fall outside voluntary transparency efforts undertaken by tech companies.[151]

To be sure, when expression is removed from one of the major online platforms, the author may find out and have a chance to appeal the decision depending on the platform's rules. But if legitimate expression is included

143 Llansó, *supra* note 123.

144 *Id.*

145 *Id.*

146 Shackford, *supra* note 63.

147 *See, e.g.*, *Counter Terrorism Internet Referral Unit*, OPEN RIGHTS GRP. WIKI, https://wiki.openrightsgroup.org/wiki/Counter_Terrorism_Internet_Referral_Unit.

148 *Id.*

149 *See id.*

150 Jane R. Bambauer & Derek E. Bambauer, *Vanished*, 18 VA. J.L. & TECH. 137, 140–41 (2013).

151 *See* ACCESS NOW, *supra* note 108, at 15.

in the shared database, its inclusion will be difficult, if not impossible, to detect. As Rebecca MacKinnon explained in her book *The Consent of the Networked*, "Once websites get on [a censorship] list, it is difficult for them to be removed, because the list itself is secret."[152]

C. Risks

Censorship creep creates serious risks for global freedom of expression.[153] Information may be removed even though it is essential for meaningful public debate and thorough reporting of the news.[154] Individuals need to speak and listen to others to govern themselves.[155] That includes the ability to participate freely in all "forms of meaning-making and mutual influence."[156] As the editorial board of the *Washington Post* wrote in response to social media companies' removal of terrorist propaganda, "Citizens of every country deserve to know what is going on in the world and what people at both ends of the spectrum think about it—however hard that is to stomach."[157]

As human rights activist Aryeh Neier (who fled the Nazis with his parents in 1939) has argued, "Freedom of speech itself serves as the best antidote to the poisonous doctrines of those who try to promote hate."[158] The public must be able to see or hear hateful or extremist views in order to interrogate and counter those views.[159] The expression of hateful or extremist ideas enables society to assert strong social norms rejecting them.[160] Those who

152 MACKINNON, *supra* note 23, at 97.

153 *See* Toor, *supra* note 39. Depending on the context, the Tech Companies may block or remove speech that would be reasonably understood as legal in most (if not all) countries; or they may block or remove speech that is legal in some countries but not in others. *See supra* note 19 and accompanying text (discussing the divergent treatment of hate speech in the United States and the EU).

154 *See* Courtney C. Radsch, *Privatizing Censorship in Fight Against Extremism Is Risk to Press Freedom*, COMM. TO PROTECT JOURNALISTS (Oct. 16, 2015), https://cpj.org/blog/2015/10/privatizing-censorship-in-fight-against-extremism-.php.

155 *See* Citron, *supra* note 6, at 191–92.

156 Jack M. Balkin, *Cultural Democracy and the First Amendment*, 110 NW. U. L. REV. 1053, 1068 (2016); *see also* Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 3 (2004); Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 PEPP. L. REV. 427, 438 (2009).

157 Editorial, *The Government Wants Social Media Sites to Take Down Terrorist Propaganda. Maybe They Shouldn't*, Wash. Post (Sept. 16, 2016), https://www.washingtonpost.com/opinions/the-government-wants-social-media-sites-to-take-down-terrorist-propaganda-maybe-they-shouldnt/2016/09/16/148d75cc-7b77-11e6-ac8e-cf8e0dd91dc7_story.html?utm_term=.ce571dc9b30e.

158 FLEMMING ROSE, THE TYRANNY OF SILENCE 85 (2014).

159 *See* Steven H. Shiffrin, *Racist Speech, Outsider Jurisprudence, and the Meaning of America*, 80 CORNELL L. REV. 43, 89 (1994).

160 C. Edwin Baker, *Autonomy and Hate Speech*, *in* EXTREME SPEECH AND DEMOCRACY, *supra* note 112, at 151.

object to hateful or terrorist ideologies can explain and justify their objections.[161]

Removal of hateful or terrorist speech on online platforms would undermine efforts designed to change people's minds.[162] For instance, Facebook launched the Online Civil Courage Initiative to counter hate speech.[163] Under the initiative, civil society groups are rewarded with advertising credits, marketing resources, and support for countering hate speech online.[164] According to Facebook's COO Sheryl Sandberg, Facebook is convinced that countering hate is a better way "to better understand and respond to the challenges of extremist speech on the internet."[165]

Similarly, Jigsaw, the Google-owned think tank, has developed a program that uses a combination of Google's advertising algorithms and YouTube's video platform to identify aspiring ISIS recruits and to dissuade them from joining the group.[166] The program places advertising alongside results for keywords and phrases commonly searched for by people attracted to ISIS.[167] The ads link to YouTube channels featuring videos that counter ISIS's brainwashing, such as testimonials from former extremists and imams denouncing ISIS's distortion of Islam.[168] As Jigsaw's Director of Research and Development Yasmin Green has explained, the company is exploring varied avenues to connect individuals with content countering violent ideologies, from ISIS to white supremacists.[169] Twitter has partnered with People Against Violent Extremism to support nongovernmental voices seeking to counter extremist ideologies.[170] The Dangerous Speech Project, led by

161 *Id.*

162 *See* Whitney v. California, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring) (arguing that the remedy for bad speech is "more speech, not enforced silence").

163 Dave Smith, *Facebook Has Launched a New Campaign Against Hate Speech*, BUS. INSIDER (Jan. 19, 2016), http://www.businessinsider.com/facebook-online-civil-courage-initiative-2016-1.

164 *Online Civil Courage Initiative*, FACEBOOK, https://www.facebook.com/pg/OnlineCivilCourage/about/?ref=page_internal (last visited Oct. 29, 2017).

165 Smith, *supra* note 163.

166 Andy Greenberg, *Google's Clever Plan to Stop Aspiring ISIS Recruits*, WIRED (Sept. 7, 2016), https://www.wired.com/2016/09/googles-clever-plan-stop-aspiring-isis-recruits/.

167 *Id.*

168 *Id.*

169 *Lawfare Podcast: Disrupting ISIS Recruitment Online*, LAWFARE (Sept. 10, 2016), https://www.lawfareblog.com/lawfare-podcast-disrupting-isis-recruitment-online.

170 *Combating Violent Extremism*, TWITTER BLOG (Feb. 5, 2016), https://blog.twitter.com/2016/combating-violent-extremism.https://blog.twitter.com/2016/combating-violent-extremism.

Susan Benesch,[171] is exploring ways for speech to counter hateful and extremist views.[172]

Even if a majority of people embracing hateful ideas may not be open to counterspeech, some may be. For instance, online discussions were instrumental to Megan Phelps-Roper's rejection of her family's hateful ideology.[173] In 2009, Phelps-Roper, a member of the Westboro Baptist Church, developed a considerable following tweeting hateful views about LGBT individuals.[174] Over time, she connected online with people who disagreed with her. Some of her online interlocutors pointed to the cruelty of her positions.[175] Phelps-Roper explained that her discussions on Twitter led her to reject the hateful views of the Westboro Baptist Church.[176] As she explained in her TED talk, her friends on Twitter:

> took the time to understand Westboro's doctrines, and in doing so, they were able to find inconsistencies I'd missed my entire life. . . . The truth is that the care shown to me by these strangers on the internet was . . . evidence that people on the other side were not the demons I'd been led to believe.[177]

Those online conversations were "life-altering" for Phelps-Roper because they helped her see that the members of her church were not the "ultimate arbiters of divine truth but flawed human beings."[178] Her life's work is now speaking out against bigotry.[179]

---

171 Colby Itkowitz, *This Professor Devotes Her Life to Countering Dangerous Speech. She Can't Ignore Donald Trump's*, WASH. POST (Oct. 24, 2016), https://www.washingtonpost.com/news/inspired-life/wp/2016/10/24/this-professor-devotes-her-life-to-countering-dangerous-speech-she-cant-ignore-donald-trumps/?utm_term=.2e1a04be21fc.

172 *See* Twitter Safety (@TwitterSafety), TWITTER (June 13, 2017, 4:54 PM), https://twitter.com/twittersafety/status/874777139895181318 ("'It's our hope that counterspeech can be a useful and tangible tool for diminishing harmful content online.' - @SusanBenesch #TSCS17.").

173 *See* Adrian Chen, *Unfollow*, NEW YORKER (Nov. 23, 2015), http://www.newyorker.com/magazine/2015/11/23/conversion-via-twitter-westboro-baptist-church-megan-phelps-roper.

174 *Id.*

175 *Id.*

176 Megan Phelps-Roper, *I Grew Up in the Westboro Baptist Church. Here's Why I Left*, TED TALK (Feb. 2017), https://www.ted.com/talks/megan_phelps_roper_i_grew_up_in_the_westboro_baptist_church_here_s_why_i_left?utm_campaign=social&utm_medium=referral&utm_source=facebook.com&utm_content=talk&utm_term=global-social%20issues#t-627390.

177 *Id.*

178 *Id.*

179 *Megan Phelps-Roper: If You're Raised to Hate, Can You Still Reverse It?*, TED RADIO HOUR (Oct. 27, 2017), https://www.npr.org/2017/10/17/560181511/megan-phelps-roper-if-youre-raised-to-hate-can-you-reverse-it. Phelps-Roper serves on Twitter's Trust and Safety Council. At the Trust and Safety Summit, she spoke with CEO Jack Dorsey about the importance of counterspeech to changing her views. Twitter Safety (@TwitterSafety), *Watch @Jack & @MeganPhelps Chat About Open & Empathetic Communication on Twitter*, TWITTER (May 2017), https://www.pscp.tv/w/1dRJZAOLjboGB.

Phelps-Roper is not alone. People have turned away from ISIS's destructive ideology thanks to counterspeech.[180] In a Brookings study entitled *The ISIS Twitter Census,* J.M. Berger and Jonathon Morgan explain that "[w]hen we segregate members of ISIS social networks, we are, to some extent, also closing off potential exit ramps."[181]

Another concern is that censorship creep will shut down discussions that allow disaffected individuals to let off steam that might prevent them from turning to violence.[182] As noted by the United Nations General Assembly in its Plan of Action to Prevent Violent Extremism, blocking online activity fuels narratives of victimization and risks further isolating disaffected individuals.[183] The risk is that aggrieved speakers will feel even more aggrieved and more inclined to act on pent-up anger.[184] Thus, removing extremist expression could "increase the speed and intensity of radicalization for those who do manage to enter the network."[185]

There are other costs beyond the realm of free expression. Removal of extremist speech may make it difficult for law enforcement to do its work. Terrorism investigations often rely on clues left in social media activity.[186] Thus, it may be difficult to investigate potential terrorism if online evidence is immediately removed.

III. Protecting Against Censorship Creep

The hate-speech agreement and industry database are here to stay—their dismantling in the near term is highly unlikely. The EU will likely continue to demand more "voluntary" changes to coerce conformity with desired speech norms. For would-be state censors, "public-private partnerships" are fruitful courses of action. They secure the adoption of governmental preferences without the burden of formal process. State actors enjoy the upside of governmental power while avoiding the messiness of political debates and

180 J.M. Berger & Jonathon Morgan, The ISIS Twitter Census 58 (2015), https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf.

181 *Id.*

182 *See* Whitney v. California, 274 U.S. 357, 375 (1927) (Brandeis, J., concurring) ("[The Framers knew] that it is hazardous to discourage thought, hope and imagination; that fear breeds repression; that repression breeds hate; that hate menaces stable government; that the path of safety lies in the opportunity to discuss freely supposed grievances and proposed remedies; and that the fitting remedy for evil counsels is good ones."); Vincent Blasi, *The Checking Value in First Amendment Theory*, 1977 Am. B. Found. Res. J. 521, 550.

183 Lucie Krahulcova, *Europol's Internet Referral Unit Risks Harming Rights and Feeding Extremism*, Access Now (June 17, 2016), https://www.accessnow.org/europols-internet-referral-unit-risks-harming-rights-isolating-extremists/.

184 Baker, *supra* note 160, at 152.

185 Berger & Morgan, *supra* note 180, at 3.

186 *See* Jenna McLaughlin, *The White House Asked Social Media Companies to Look for Terrorists. Here's Why They'd #Fail*, Intercept (Jan. 20, 2016), https://theintercept.com/2016/01/20/the-white-house-asked-social-media-companies-to-look-for-terrorists-heres-why-theyd-fail/.

judicial hearings. Private reporting processes also permit far greater censorship than law would allow. Platforms' definitions of hate speech can be stretched to cover all forms of disfavored content including political dissent. EU regulators have no reason to depart from this approach given the success of extralegal efforts.

Laws designed to check the potential for censorship creep are unlikely. Rebecca MacKinnon put it well: "[C]orporate collusion in government surveillance and censorship is unlikely to be solved by the passage and enforcement of laws, even by the most well-intentioned and democratic of governments."[187] Ultimately, Silicon Valley may be our best protection against censorship creep.

As this Part outlines, tech companies should adopt special policies and procedures to protect against governmental overreach. The goal would be to enhance the clarity, accountability, and transparency of censorship efforts. Ombudsmen should be hired to check governmental attempts to suppress newsworthy content. These proposals are offered in hopes that executives might view them as an effective way to contain EU censorship pressure and to convey their support of users' expression.

A. Definitions

Clarity in the definition, meaning, and application of the terms "hate speech" and "terrorist material" would help contain censorship creep. In 2011, Helen Norton and I argued that when companies ban hate speech, they ought to explain precisely what the term means and why it has been banned.[188] In our view, users needed this information to understand their rights and responsibilities when using platforms.[189] Definitional clarity serves another goal: preventing private hate-speech bans from being leveraged to silence legitimate expression.

Professor Norton and I laid out potential definitions for hate speech based on principles of U.S. tort and civil rights law.[190] For instance, hate speech could be defined as that which threatens or encourages violence against traditionally subordinated groups (or specific members of groups).[191] It could be defined as speech that intentionally inflicts emotional distress—expression that is individually targeted, especially threatening or humiliating, repeated, and reliant on sensitive or outrageous material.[192] It could be defined as speech that would rise to the level of

187 MACKINNON, *supra* note 23, at 175.

188 Citron & Norton, *supra* note 7, at 1459. In hopes that companies might find our suggestions helpful, we presented our article at a meeting of the Anti-Cyber Hate Working Group held at Stanford Law School in 2011.

189 *See id.* at 1457–58.

190 *See id.* at 1460–68.

191 *See id.* at 1461–62.

192 *Id.* at 1463.

actionable workplace harassment under civil rights laws.[193] Those definitions, however, may be too narrow to satisfy EU regulators.

The Tech Companies could look to international human rights law for guidance.[194] The problem of ambiguity, however, would remain because human rights law contains exceptionally flexible standards.[195] Another possibility is to consult the work of the Council of Europe's Secretary General who is devising "common European standards [for hate speech and terrorist material] to better protect freedom of expression online."[196] Those efforts hopefully will provide definitions that curtail the malleability of those terms.[197]

Human rights groups have expressed serious reservations about the hate-speech agreement because hate speech is not well defined. EU Justice Commissioner Jourová responded to their objections by pointing to the EU Framework Decision, European Court of Human Rights jurisprudence, and the law of member states.[198] At the very least, companies should hold the European Commission to that position and look to those sources in devising their own definitions for hate speech and extremist material.

---

193 *See id.* at 1464–65.

194 *See* Scott Craig & Emma Llansó, *Pressuring Platforms to Censor Content Is Wrong Approach to Combatting Terrorism*, CDT BLOG (Nov. 5, 2015), https://cdt.org/blog/pressuring-platforms-to-censor-content-is-wrong-approach-to-combatting-terrorism/ (arguing that when government seeks to police speech, notably extremism, through terms of service, those requests "should be grounded in legal frameworks rooted in international human rights" rather than terms of service).

195 *See* Rose, *supra* note 158, at 150–51 (explaining that international law does not embrace a universally accepted definition of hatred while member states generally define hate speech as utterances expressing hatred or antipathy towards other social groups, a standard that makes line drawing difficult if not impossible). As Floyd Abrams explains in his most recent book *The Soul of the First Amendment*, the European Court of Human Rights has upheld hate-speech convictions involving criticism of politicians and bigoted views expressed by politicians. ABRAMS, *supra* note 21, at 41–42.

196 *Council of Europe Secretary General Concerned About Internet Censorship: Rules for Blocking and Removal of Illegal Content Must Be Transparent and Proportionate*, COUNCIL OF EUR. (June 1, 2016), https://www.coe.int/en/web/tbilisi/-/council-of-europe-secretary-general-concerned-about-internet-censorship-rules-for-blocking-and-removal-of-illegal-content-must-be-transparent-and-prop.

197 *See id.* The Secretary General remarked that while governments have an obligation to combat the promotion of terrorism, hate speech, and other illegal online content,

> I am concerned that some states are not clearly defining what constitutes illegal content. Decisions are often delegated to authorities who are given a wide margin for interpreting content, potentially to the detriment of free expression. On the basis of this study we will take a constructive approach and develop common European standards to better protect freedom of expression online.

Id.

198 Letter from Vera Jourová, EU Justice Comm'r, to Jens-Henrik Jeppesen, Sec'y Gen. of the Council of Eur. (June 21, 2016). Commissioner Jourová has not adhered to this definition in talking about the hate-speech agreement. As Part I explained, she has suggested that fake news, online radicalization, and terrorist propaganda should be removed pursuant to the agreement.

What about the industry database of banned terrorist content? The laws of EU member states provide little guidance because they are exceptionally broad.[199] The Tech Companies have said that they only intend to share hashes of "the most extreme and egregious terrorist images and videos"—"content most likely to violate all of our respective companies' content policies."[200] Concerns about censorship creep can be tackled with policies that explain precisely what the phrase "the most extreme and egregious terrorist images and videos" means. Policies should provide specific examples of content deserving of that designation. Such clarity, explanations, and examples can help prevent the gradual broadening of the standards governing what content is included in the database.

As the Tech Companies work on their definitions of hate speech and extremist material, they might consider including human rights groups as well as academics in their efforts.[201] Building on the work of the Anti-Cyberhate Working Group[202] and the Global Network Initiative,[203] a multistakeholder group could be established to help companies articulate what constitutes hate speech or terrorist material. Civil liberties groups have argued for a role in helping companies understand "the various meanings given to 'violent extremism' and related concepts, and the potential impact of ambiguity in this area on the promotion and protection of human rights."[204]

---

199 National laws differ as to what terrorist content is deemed unlawful and in what contexts; some of those laws are so broad that they have been used to jail journalists, bloggers, and human rights defenders. *See* Letter from Judith Lichtenberg, Exec. Dir., Global Network Initiative, to David Kaye, UN Special Rapporteur, Office of the UN High Comm'r for Human Rights (Jan. 29, 2016), http://www.ohchr.org/Documents/Issues/Expression/PrivateSector/GlobalNetworkInitiative.pdf.

200 Clark, *supra* note 65.

201 *See* EUROPEAN DIG. RIGHTS, INPUT ON HUMAN RIGHTS AND PREVENTING AND COUNTERING VIOLENT EXTREMISM (Mar. 18, 2016), https://edri.org/files/2016-UN-consultation.pdf (criticizing the Europol Internet Referral Unit and other authorities that pressure companies to remove terrorist or extremist content and thus allow states to regulate and stifle online speech indirectly).

202 *See* Klonick, *supra* note 6, at 67. For instance, the Anti-Cyber Hate Working Group, of which I was a member, met annually, first at Stanford in 2012 and then in subsequent years at tech companies. *See* Citron, *supra* note 6, at 232. The goal was to devise best practices for dealing with online abuse and hate speech. *See ADL Releases "Best Practices" for Challenging Cyberhate*, ADL (Sept. 23, 2014), https://www.adl.org/news/press-releases/adl-releases-best-practices-for-challenging-cyberhate.

203 The Global Network Initiative is a multistakeholder group made up of forty organizations: technology companies, including Facebook, Google, Microsoft, and Yahoo!; human rights groups, such as the Center for Democracy and Technology, Human Rights Watch, and the Committee to Protect Journalists; free speech and privacy experts like Deirdre Mulligan and Rebecca MacKinnon; and others devoted to forging a common approach to free expression and privacy online. *See Participants*, GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/participants/index.php (last visited Nov. 7, 2017).

204 *UN HRC: Resolution on "Violent Extremism" Undermines Clarity*, ARTICLE 19 (Oct. 8, 2015), https://www.article19.org/resources.php/resource/38133/en/un-hrc:-resolution-on-%E2%80%9Cviolent-extremism%E2%80%9D-undermines-clarity.

Those definitions, designed for content moderators, should be shared publicly so governments can understand the limits of efforts to remove speech under TOS agreements. With this knowledge, governments might be less emboldened to push companies to broaden hate-speech or terrorism policies beyond recognition. Some might even reconsider trying to silence unpopular but protected expression.

B. Robust Accountability

Removal requests made by state authorities (or nongovernmental organizations acting on the state's behalf) should be subject to rigorous review. As a start, government officials should be required to identify themselves when reporting content for TOS violations. Online platforms must know that they are dealing with governmental authorities or their surrogates. There should be a separate reporting channel for government authorities and any organization working on a state's behalf. For instance, Twitter has "intake channels [dedicated] for law enforcement and other authorized reporters" to file "legal requests."[205] That is a good start. All removal requests—"legal requests" *and* TOS reports—should proceed through that channel.

What about the review process itself? As Kate Klonick has documented in her groundbreaking work,[206] Facebook has extensive internal review processes with detailed instructions about what content is banned and under what circumstances.[207] In-house employees or outside companies like Sutherland and Deloitte make decisions about content moderation on the company's behalf.[208] To ensure that terms of service are enforced uniformly, moderators receive extensive training on the rules as well as on potential cultural biases and emotional reactions.[209]

---

205 *Removal Requests*, TWITTER, https://transparency.twitter.com/en/removal-requests.html (last visited Nov. 16, 2017). The separate reporting channel applies to "legal requests," which includes court orders and formal requests from government agencies and law enforcement. *Id.*

206 Klonick, *supra* note 6, at 43–57 (exploring the normative significance of a complex knit of rules devised by platforms to govern the content moderation process).

207 In 2012, a version of Facebook's instruction manual for reviewers was leaked online. Danielle Citron, *Actualizing Digital Citizenship with Transparent TOS Policies: Facebook Style*, CONCURRING OPINIONS (Mar. 16, 2012), https://concurringopinions.com/archives/2012/03/actualizing-digital-citizenship-with-transparent-tos-policies-facebooks-leaked-policies.html/comment-page-1. As Kate Klonick shows, since that time, Facebook has developed even more extensive and detailed instructions for individuals reviewing abuse complaints. Klonick, *supra* note 6, at 41; *see* Julia Angwin, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children*, PROPUBLICA (June 28, 2017), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms. Those instructions could be revised with special instructions for removal requests made by government actors.

208 *See* Klonick, *supra* note 6, at 49 (discussing the different tiers of moderators in the escalation process).

209 *See id.* at 51–53.

Other online platforms approach content moderation in a similarly careful way. As part of that process, all companies should ensure that government requests are viewed through a special lens. When government authorities seek to suppress speech under terms of service, content moderators should view requests with a presumption against removal, or at least a healthy dose of skepticism.[210] Content moderators should receive training about censorship creep, including past and present governmental efforts to silence critics. Training should focus on how to distinguish banned hate speech or terrorist material from newsworthy content. This is not an easy task but crucial nonetheless.

Decisions related to government requests should be accompanied by an explanation—decisionmakers who have to articulate their reasons are likely to think more carefully about their decisions.[211] When a moderator decides to grant a government request for removal based on a TOS violation, that decision should automatically pass through a second layer of review. Individuals whose speech is removed should be notified about the removal and given a chance to appeal.[212]

The hate-speech agreement's requirement that the Tech Companies resolve reports within twenty-four hours will make it difficult and expensive to implement additional layers of review. With well-trained moderators, detailed policies, and a presumption against removal, timeliness may become less of a problem.

The actual decisions, of course, may provoke the ire of the European Commission. The European Commission will not be pleased if removal requests are not granted, much as the recent review suggested. That tension will remain until the Tech Companies and European users can convey to the Commissioners and lawmakers their concerns about free expression, much as Apple did in the recent encryption debate in the United States. If human rights groups were given the ability to act as advisers and watchdogs, then they could help explain to the public the real risks of censorship creep.

What about the industry database of banned violent extremist content? As the previous Part explored, the industry database could become a blacklist. Once included in the database, content may never end up appearing on any of the major platforms. Inclusion in the database has greater potential to systematically silence speech than individual requests to remove content under terms of service. As a result, strong protections are essential to prevent governments from coopting the database.

---

210 *Written Comments of Article 19 on the Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/32/28 (May 11, 2016), http://www.ohchr.org/Documents/Issues/Expression/PrivateSector/Article19.pdf.

211 Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1302 (2008).

212 The dominant online platforms do just that for decisions resulting in the removal of speech. Many provide an appeals-like process so users can present their objections.

One option is for companies to adopt a blanket rule that governments cannot contribute hashes to the database. As Emma Llansó argues, the Tech Companies should "[c]learly and unequivocally state that under no circumstances will [they] accept a contribution of a hash by someone acting on behalf of a government."[213] To avoid efforts to undermine that ban, they could announce any effort or request from a government to undermine that policy.[214] Another alternative is for companies to subject a government request to several layers of review and to condition the submission on the approval of senior staff.[215] Rigorous accountability can help prevent systematizing censorship of legitimate speech at the behest of European governments and other nations.

C. Meaningful Transparency

Another check on censorship creep is for companies to provide detailed reports on governmental efforts to censor hate speech and extremist material through informal measures. Transparency reports enable public conversation about censorship. In turn, European users can contact lawmakers with concerns about attempts to use tech companies as censorship proxies. The more users find out about companies' efforts to protect their fundamental freedoms, the more trust users will have in online providers.[216] Human rights advocates can call attention to concerns about censorship creep. Ultimately, transparency reports can generate "productive discussion about the appropriate use and limits of [state] authority."[217]

The Electronic Frontier Foundation began tracking the extent to which companies were being transparent about how often they are blocking or removing content or accounts in 2015. At that time, Google was the only tech company providing transparency reports.[218] Since then, many more platforms have endeavored to provide some transparency about government requests to suppress speech. Reviewing the successful features of current efforts can help formulate a path forward.

Twitter has been hailed for its transparency efforts, and rightfully so. The company's 2016 Transparency Report details the number of legal requests for content removal based on country.[219] Crucially, and uniquely, it

213 Llansó, *supra* note 123.

214 *See id.*

215 *See id.*

216 For work on the importance of trust in companies' respect for privacy and expression, see Neil Richards and Woodrow Hartzog's important scholarship including *Taking Trust Seriously in Privacy Law*, 19 STAN. TECH. L. REV. 431 (2016), and *Privacy's Trust Gap: A Review*, 126 YALE L.J. 1180 (2017) (reviewing FINN BRUNTON & HELEN NISSENBAUM, OBFUSCATION: A USER'S GUIDE FOR PRIVACY AND PROTEST (2015)).

217 Liane Lovitt, *Why Transparency Reports Matter Now More than Ever*, MEDIUM (May 13, 2016), https://medium.com/inflection-points/why-transparency-reports-matter-now-more-than-ever-9fb6ebe733fa#.yhossoi1q.

218 Thanks to Daphne Keller for pointing this out to me.

219 *See Removal Requests*, *supra* note 205.

discloses the number of government requests seeking removal of terrorism content for TOS violations.[220] Twitter is "actively working to expand" its reporting of all "known, non-legal government TOS requests . . . through [its] standard . . . intake channels, such as requests to remove impersonating accounts and other content that violates our Rules against abuse."[221] Although Facebook's 2016 Transparency Report provided important information to users, it did not include the number of government requests seeking removal of content based on its terms of service and community standards.[222]

Much as Twitter has done for terrorist content and expects to do far more of in the future, corporate transparency reports should detail the number, subject matter, and results of *all* government requests to remove content for TOS violations.[223] If governments are allowed to request the addition of hashes to the industry database, transparency reports should include details about those requests. Although transparency cannot solve the problem of censorship creep, it can help contain it, especially if strong standards and robust accountability procedures are adopted.

D. Ombudsmen

An acute problem related to censorship creep is its potential to suppress newsworthy content. Government removal requests may seek to remove terrorist or hateful content whose publication is in the public's interest. To address this concern, companies should consider hiring or consulting

220 *Government TOS Reports*, TWITTER, https://transparency.twitter.com/en/gov-tos-reports.html (last visited Nov. 16, 2017) (reporting that in the six-month period from July 2016 to December 2016, Twitter received 716 reports from governments related to 5929 accounts and that 85% were removed for TOS violations related to violent extremism).

221 *Id.* In its 2016 Transparency Report, Microsoft revealed the number of requests made by countries for the removal of content based on local laws or TOS violations and the percentage that were granted. *Content Removal Requests Report*, MICROSOFT, https://www.microsoft.com/en-us/about/corporate-responsibility/crrr (last visited Nov. 16, 2017). It did not, however, tease out formal from informal requests as Twitter did.

222 *See Government Requests Report: FAQs*, FACEBOOK, https://govtrequests.facebook.com/faq/. But Facebook's 2016 Transparency Report did helpfully provide examples of governmental requests to remove content under local laws and its response. For instance, France asked Facebook to remove a photograph depicting victims of the terrorist attack at the Bataclan concert hall. As the report explained, the photo did not violate Facebook's community standards, but because it violated French law, access to the photo was blocked in France. *Id.*

223 *See* FREEDOM ONLINE COALITION, SUBMISSION TO UN SPECIAL RAPPORTEUR DAVID KAYE: STUDY ON FREEDOM OF EXPRESSION AND THE PRIVATE SECTOR IN THE DIGITAL AGE (2016), https://www.freedomonlinecoalition.com/wp-content/uploads/2016/02/FOC-WG3-Submission_ICT-Sector-Report.pdf. Some countries like China do not allow transparency reporting. *See* RANKING DIGITAL RIGHTS, SUBMISSION TO UN SPECIAL RAPPORTEUR DAVID KAYE: STUDY ON FREEDOM OF EXPRESSION AND THE PRIVATE SECTOR IN THE DIGITAL AGE (2016), http://www.ohchr.org/Documents/Issues/Expression/PrivateSector/RankingDigitalRightsAndNewAmerica.pdf.

ombudsmen whose life's work is the newsgathering process.[224] Ombudsmen, who are also known as public editors, work to "protect press freedom" and to promote "high-quality journalism."[225] Their role is to "act[ ] in the best interests of news consumers."[226]

Ombudsmen should have a special role in assessing government removal requests made through informal channels like terms of service or the industry database. They can help identify requests that would remove material that is important for public debate and knowledge. Then too, because the industry database raises special concerns about the suppression of expression, the ombudsman could review all contributions to the database with the public interest in mind.[227]

E. Geographically Tailored TOS

Might tech companies limit the impact of EU pressure by refusing to incorporate EU norms into terms of service? Companies could tailor terms of service and community guidelines to specific countries and regions. This would map speech norms onto countries and regions rather than imposing a one-size-fits-all model across the globe.

Policies that balkanize online practices do run the risk of endangering the free flow of information. Consider the impact of laws that require companies to store data collected from a country's citizens within its borders. Data localization laws, as they are known, bring all online interactions under the control of local governments, which is especially troubling in the case of authoritarian regimes.[228] Iran's data localization law is viewed as a tool for the government to stifle dissent.[229]

Tailoring terms of service by country or region does risk curtailing free expression. But it would also have an important upside. It would prevent the removal or blocking of speech in more speech-protective countries while allowing the removal or blocking of speech in countries with more restrictive speech norms.[230]

---

224 *See About ONO*, ONO, http://newsombudsmen.org/about-ono (last visited Jan. 8, 2018).

225 *Id.*

226 *Id.*

227 The database should be accessible to and analyzed by independent experts familiar with free expression and policy. *See* Jeppesen, *supra* note 108.

228 *See* ANUPAM CHANDER & UYÊN P. LÊ, BREAKING THE WEB: DATA LOCALIZATION VS. THE GLOBAL INTERNET 46–47 (2014), https://pdfs.semanticscholar.org/0b70/1f2601ff39bc338dc351e5c8a6a40f98f304.pdf.

229 *Id.* at 47.

230 My analysis in this Article focuses on content hosts, the top layer of the internet. The concerns that I raise in this piece take on even greater significance the more power over speech that companies possess and the fewer the options available to consumers. This is certainly true of internet service providers and domain name registrars among other layers in the internet stack. This is true of companies like Cloudflare whose services are indispensable for remaining online. *See* Kate Klonick, Opinion, *The Terrifying Power of Internet Censors*, N.Y. TIMES (Sept. 13, 2017), https://www.nytimes.com/2017/09/13/opin

CONCLUSION

Silicon Valley has a long history of embracing American-style speech norms. For more than a decade, when social media platforms admitted exceptions to the notion that information should be free, they were careful to ensure that those exceptions remained narrow. This was true whether the issue was impersonation, stalking, or nonconsensual pornography. Free expression has long been embedded in tech companies' corporate culture.

In the wake of recent terror attacks and the resurgence of hate groups, European regulators have pressured tech companies to change their speech rules and practices to remove extremist material appearing on their platforms. Some of the resulting changes may do some good. After all, an industry database that flags for removal videos containing "kill lists" or bomb instructions may ultimately prevent violence. The hate-speech agreement could lead to the rapid removal of posts calling for the death of religious minorities, which, if left up, could inspire physical violence on members of those groups.

But the changes to private rules and practices should be understood and evaluated for what they are: compelled conformity with European speech norms. Despite the protestations of EU regulators, they are neither voluntary nor the product of meaningful public-private partnerships. Instead, they are the result of government coercion occurring outside the rule of law. What is different about the pressure from states now is that it has brought about changes that risk worldwide censorship creep. Because governments are using terms of service to achieve their ends, the resulting suppression of speech will be global.

This Article offers potential safeguards to prevent censorship creep. Companies can and should adopt prophylactic protections designed to manage extralegal pressure for the good of free expression. Greater clarity, accountability, transparency, and oversight would help check EU efforts to censor speech on a global scale. As companies assess these suggestions, a multistakeholder approach could provide crucial expertise to tech companies as they develop processes to resist censorship creep while combating the real perils of hate speech and extremist material. A "solution—conceived of and imposed primarily by some combination of government and corporations—will likely . . . be inadequate to deal with the complex issues raised by the problem of extremist use of social media."[231] It should include experts in extremism and hate, victims of extremist violence, scholars of political dis-

ion/cloudflare-daily-stormer-charlottesville.html?mcubz=3&_r=0 (discussing Cloudflare's refusal to provide its security services to neo-Nazi website Daily Stormer in the wake of the armed white supremacist rally in Charlottesville and murder of counterprotestor Heather Heyer). Neil Richards and I will be exploring concerns about power and digital speech in an article entitled *Essential Preconditions for Free Expression in the Digital Age*, 95 WASH. U. L. REV. (forthcoming 2018). Thanks to Michael Nelson for discussing these concerns with me at the CATO Free Expression symposium.

231 BERGER & MORGAN, *supra* note 180, at 61.

sent, and civil rights and civil liberties organizations.[232] It would deepen the efforts of companies, such as Facebook, which is bringing counterterrorism expertise in house,[233] and Twitter, whose Trust and Safety Council provides advice on potential proposals concerning online abuse.[234]

As Apple's struggle with the U.S. government over encryption illustrated and as Silicon Valley's unanimous support for that stand reaffirmed, tech companies enjoy public support when they defend fundamental freedoms. The suggestions outlined in this piece thus may be positively received.

232 *Id.*

233 Curt Mills, *Facebook Is Looking for a Counterterrorism Analyst*, U.S. NEWS & WORLD REP. (Nov. 14, 2016), https://www.usnews.com/news/national-news/articles/2016-11-14/facebook-is-looking-for-a-counterterrorism-analyst.

234 *The Twitter Trust and Safety Council*, TWITTER, https://about.twitter.com/safety/council (last visited Jan. 26, 2018).