2019

The Marketplace of Ideas Online

Dawn C. Nunziato
George Washington University Law School, dnunziato@law.gwu.edu

Recommended Citation

Nunziato, Dawn Carla, The Marketplace of Ideas Online (2019). The Marketplace of Ideas Online, 94 Notre Dame L. Rev. 1519-1584 (2019).; GWU Law School Public Law Research Paper No. 2019-36; GWU Legal Studies Research Paper No. 2019-36. Available at SSRN: https://ssrn.com/abstract=3405381 or http://dx.doi.org/10.2139/ssrn.3405381

THE MARKETPLACE OF IDEAS ONLINE

*Dawn Carla Nunziato**

* William Wallace Kirkpatrick Research Professor of Law, The George Washington University Law School; Codirector, Global Internet Freedom Project. I am grateful to the editors of the *Notre Dame Law Review* and the participants in the *Notre Dame Law Review* Symposium on *Contemporary Free Speech: The Marketplace of Ideas a Century Later*, especially Alexander Tsesis and Fred Schauer, and to my colleagues Chip Lupu and Todd Peterson, for their helpful comments on this Article. I am also grateful to Alexia Khella and Ken Rodriguez for providing excellent research and library assistance in connection with this Article, to Kierre Hannon for excellent administrative assistance, and to Dean Blake Morant for financial support of my research.

INTRODUCTION

One hundred years ago, in the 1919 case of *Abrams v. United States*, Justice Oliver Wendell Holmes ushered into existence modern First Amendment jurisprudence by introducing the "free trade in ideas" model of free speech.[1] According to this model, the ultimate good is reached by allowing speakers to engage in the free trade in ideas—free of government intervention in the way of regulation, censorship, or punishment. Ideas must be allowed to compete freely in an unregulated market, and the best ideas will ultimately get accepted by competing with others in this marketplace. As such, government intervention is unnecessary and counterproductive. Thus, instead of punishing the speakers in *Abrams*—for criticizing the government's attempts to crush the Russian Revolution and calling for American workers to strike[2]—the government should have taken a hands-off approach and allowed these ideas to compete (and lose) in the marketplace of ideas.[3]

The characteristics of our marketplace(s) of ideas have changed dramatically since 1919, when the Russian immigrants in *Abrams* threw their leaflets from the fourth floor window of a hat factory in lower Manhattan in an effort to widely disseminate their ideas.[4] Russians are still players in our marketplace of ideas, but today's marketplace suffers from uniquely modern and challenging problems—such as rampant interference in the form of Russian troll farms mass-producing tweets and other widely shared content on social media with the intent and the effect of sabotaging U.S. elections.[5] In addi-

1 Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting).
2 *Id.* at 617–19 (majority opinion).
3 *Id.* at 630 (Holmes, J., dissenting).
4 *Id.* at 617–18 (majority opinion).
5 *See* OFFICE OF THE DIR. OF NAT'L INTELLIGENCE, NAT'L INTELLIGENCE COUNCIL, INTELLIGENCE COMMUNITY ASSESSMENT 2017-01D, ASSESSING RUSSIAN ACTIVITIES AND INTENTIONS

tion to the widespread dissemination of false political content from both foreign and domestic sources, today's online marketplace of ideas is besieged by the increased polarization and siloing of thought and opinion, which renders Holmes's prescribed remedy for harmful speech—counterspeech—increasingly ineffective.[6]

In the past two years we have seen a variety of efforts, both in the United States and across the globe, by governments and by online platform providers themselves, to address the problems, distortions, and imperfections in the online marketplace. Because online platforms like Facebook and Twitter play such a dominant role in the online marketplace of ideas—and the modern marketplace of ideas generally—it is worthwhile to focus specifically on how these platforms are being regulated, as well as how they are regulating themselves. While the United States has essentially taken a hands-off approach to regulating online platforms, the European Union has assumed a relatively aggressive regulatory approach.[7] The European Union, as well as several European countries, have generally implemented speech regulations to hold platforms liable for failing to police their sites, and have recently imposed sweeping regulations on such platforms. And, in their efforts to comply with such regulations, online platforms like Facebook and Twitter may end up implementing these European regulations in ways that affect what U.S. audiences can access online—since it is often difficult for platforms to implement national regulations in a geographically targeted manner with no spillover beyond the regulating nation's borders.[8] Accordingly, it is worthwhile to examine these international efforts in some detail. The European Union and European countries have recently undertaken sweeping efforts to remedy perceived imperfections in the marketplace,[9] including by requiring online platforms to rapidly remove a wide swath of harmful content.[10] Among European nations, Germany has led the way by enacting drastic legislation requiring social media sites like Facebook and Twitter to remove false news, defamatory hate speech, and other unlawful content within twenty-four hours of receiving notice of the same, upon pain of multi-

IN RECENT US ELECTIONS 2 (2017), https://www.dni.gov/files/documents/ICA_2017_01.pdf.

6 *See, e.g.*, Cristina Maza, *Florida Shooting: Russian Bots Flooded the Internet with Propaganda About Parkland Massacre*, NEWSWEEK (Feb. 16, 2018), https://www.newsweek.com/florida-shooting-russian-bots-twitter-809000 (reporting on Russian-linked bots tweeting about the Parkland shooting and gun reform).

7 *See infra* Section III.A (comparing Communications Decency Act section 230 and the European Union approach to online intermediary liability).

8 *See infra* Section III.A.

9 *See, e.g.*, Joanna Plucinska, *Macron Proposes New Law Against Fake News*, POLITICO (Jan. 3, 2018), https://www.politico.eu/article/macron-proposes-new-law-against-fake-news (reporting on French President Emmanuel Macron's proposal for "new rules cracking down on fake news").

10 *See infra* text accompanying notes 60–72.

million-euro fines.[11] Other European countries are considering following suit.

In addition to government regulation by the European Union and by European governments, the online platforms themselves are undertaking self-regulatory measures with respect to content accessible by U.S. audiences (partly in an effort to forestall U.S. government regulation).[12] Although such self-regulatory efforts are not governed by the First Amendment, they are nonetheless inspired by First Amendment values.[13] The leading social media companies have adopted several measures to attempt to address problems in the online marketplace of ideas, including by enabling the flagging of false news for verification by independent third-party fact-checkers,[14] commissioning the development of counterspeech in response to false news, providing contextual information about purveyors of news-related posts, and removing fake sites and purveyors of false news from their platforms.[15]

Although the United States has largely taken a hands-off approach to regulating online platforms, in the wake of the severe problems besieging the platforms in the context of the 2016 presidential election and thereafter, U.S. legislators have recently sought to hold the online platforms responsible for such problems. In addition to extensive legislative hearings during which legislators have sought to hold the companies to account for such problems,[16] legislators have recently proposed new laws to attempt to remedy such problems. In particular, Congress recently proposed the Honest Ads

11 *See Overview of the NetzDG Network Enforcement Law*, CTR. FOR DEMOCRACY & TECH. (July 17, 2017), https://cdt.org/insight/overview-of-the-netzdg-network-enforcement-law.

12 *See, e.g.*, *Germany: Flawed Social Media Law: NetzDG Is Wrong Response to Online Abuse*, HUM. RTS. WATCH (Feb. 14, 2018), https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law [hereinafter *Flawed Social Media Law*] ("At least three countries—Russia, Singapore, and the Philippines—have directly cited the German law as a positive example as they contemplate or propose legislation to remove 'illegal' content online.").

13 As Marvin Ammori explains, those at the helm of such companies report that they are inspired by and committed to First Amendment values in general and the marketplace of ideas model in particular. Marvin Ammori, *The "New"* New York Times*: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2262 (2014).

14 *See* Tessa Lyons, *Hard Questions: How Is Facebook's Fact-Checking Program Working?*, FACEBOOK NEWSROOM (June 14, 2018), https://newsroom.fb.com/news/2018/06/hard-questions-fact-checking (describing the steps taken in the fact-checking process, including identifying potentially false stories, reviewing the story, and action taken against false stories); Sydney Schaedel, *How to Flag Fake News on Facebook*, FACTCHECK.ORG (July 6, 2017), https://www.factcheck.org/2017/07/flag-fake-news-facebook.

15 *See* Sarah Frier, *Facebook Has Removed More than 800 U.S. Accounts Spreading Fake News*, TIME (Oct. 11, 2018), http://time.com/5422546/facebook-removes-800-fake-news-accounts/ (reporting that Facebook has "removed 559 pages and 251 accounts that were coordinating the spread of misinformation and spam in the U.S."); *see also* Sheera Frenkel, *Facebook to Remove Misinformation that Leads to Violence*, N.Y. TIMES (July 18, 2018), https://www.nytimes.com/2018/07/18/technology/facebook-to-remove-misinformation-that-leads-to-violence.html (reporting that Facebook will remove *content* as misinformation if it "could lead to people being physically harmed").

16 *See, e.g.*, *Transcript of Mark Zuckerberg's Senate Hearing*, WASH. POST (Apr. 10, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-

Act in an effort to limit foreign interference in the online marketplace of ideas and to mandate the disclosure of information regarding the source of political advertisements on social media.[17]

Finally, in the United States, victims and targets of some of the problems besieging the online marketplace of ideas—including false news, conspiracy theories, and hoaxes—are increasingly turning to defamation law in an effort to hold the purveyors to account for the harms resulting from such online content.[18]

This Article surveys the severe problems in today's online marketplace of ideas and the efforts that regulators—and the online platforms themselves—have recently adopted in an attempt to address such problems. In Part I, this Article examines the historical foundations of the "marketplace of ideas" model, as articulated in Holmes's early opinions, as well as the Court's eventual adoption of the marketplace model and, with it, the adoption of counterspeech, instead of censorship, as the default response to harmful speech. Part II then examines the scope and extent of the problems besieging the modern online marketplace of ideas, focusing on problems that have arisen especially in the context of the 2016 U.S. presidential election and thereafter on social media platforms like Facebook and Twitter. In Section III.A, this Article examines the sweeping regulatory efforts recently adopted by the European Union and by Germany in particular, and the ways in which the online platforms are striving to implement such regulations. In Sections III.B and III.C, the Article turns to an analysis of the self-regulatory efforts undertaken by leading social media platforms Facebook and Twitter, the likely efficacy of such measures in addressing the problems besieging the online marketplace of ideas, and the extent to which such measures are consistent with First Amendment values. In Section III.D, the Article examines the constitutionality and the likely efficacy of the recently proposed Honest Ads Act. In Section III.E, the Article examines the extent to which the defamation lawsuits brought by victims of false news, conspiracy theories, and online hoaxes are consistent with the First Amendment. A brief conclusion follows.

I. The Historical Origins of the Marketplace of Ideas

The Supreme Court's marketplace of ideas model of the First Amendment has its roots in Holmes's "free trade in ideas"[19] formulation, which

zuckerbergs-senate-hearing/ (transcribing Senate hearing of Facebook CEO Mark Zuckerberg).

17 Honest Ads Act, S. 1989, 115th Cong. (2017).

18 *E.g.*, Elizabeth Williamson, *Alex Jones, Pursued over Infowars Falsehoods, Faces a Legal Crossroads*, N.Y. Times (July 31, 2018), https://www.nytimes.com/2018/07/31/us/politics/alex-jones-defamation-suit-sandy-hook.html.

19 Note that Holmes—in his judicial opinions and in his writings—did not use the phrase "marketplace of ideas" and the Court itself did not adopt this formulation until 1953 with Justice William O. Douglas's concurrence in *United States v. Rumely*, 345 U.S. 41, 56 (1953) (Douglas, J., concurring). *See* Christoph Bezemek, *The Epistemic Neutrality of the*

places primacy on the individual speech decisions of market actors unfettered by government regulation, as articulated by Holmes in several dissenting opinions in the early twentieth century.[20] In his first significant First Amendment opinion, Holmes dissented from a decision upholding the prosecution of five individuals for encouraging resistance to the United States' action in World War I.[21] In *Abrams v. United States*,[22] Holmes championed a robust marketplace of ideas constituted by the free speech decisions of individuals like the *Abrams* defendants, free of state censorship:

> [T]he ultimate good desired is better reached by *free trade in ideas*—that the best test of truth is the power of the thought to get itself accepted in the competition of the market . . . . That at any rate is the theory of our Constitution. It is an experiment, as all life is an experiment. . . . While that experiment is part of our system I think that we should be eternally vigilant against attempts to check the expression of opinions that we loathe and believe to be fraught with death, unless they so imminently threaten immediate interference with the lawful and pressing purposes of the law that an immediate check is required to save the country.[23]

Influenced by his intellectual predecessors John Milton[24] and John Stuart Mill,[25] as well as by his contemporaries Judge Learned Hand, Harold

---

"Marketplace of Ideas": Milton, Mill, Brandeis, and Holmes on Falsehood and Freedom of Speech, 14 FIRST AMEND. L. REV. 159, 172–73 (2015); Thomas W. Joo, *The Worst Test of Truth: The "Marketplace of Ideas" as Faulty Metaphor*, 89 TUL. L. REV. 383, 390 (2014).

20 Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting); *see also* Gitlow v. New York, 268 U.S. 652, 672–73 (1925) (Holmes, J., dissenting).

21 *Abrams*, 250 U.S. at 624.

22 *Abrams* involved the prosecution of five Russians for violating the Espionage Act for encouraging resistance to the United States in the war against Germany in World War I. *Id.* at 616–19 (majority opinion).

23 *Id.* at 630 (Holmes, J., dissenting) (emphasis added). Holmes adopts in this passage a skeptical theory of ultimate truth—one in which ultimate "truth" cannot be known or identified, but merely consists in whatever emerges as the winner in the competition in the marketplace of ideas. *Id.*

24 John Milton, while primarily renowned as one of the greatest English poets, was also an outspoken opponent of state censorship. In response to a statute passed by the English Parliament that prohibited publication of any book without government approval, Milton argued forcefully in favor of freedom of expression against government censorship. In this representative passage from his *Areopagitica*, Milton poetically set forth the foundation for the marketplace of ideas conception of freedom of expression, which Holmes later relied on:

> And though all the winds of doctrine were let loose to play upon the earth, so Truth be in the field, we do injuriously by licensing and prohibiting to misdoubt her strength. Let her and Falsehood grapple; who ever knew Truth put to the worse, in a free and open encounter?

JOHN MILTON, AREOPAGITICA 58 (Richard C. Jebb ed., Cambridge Univ. Press 1918) (1644).

25 John Stuart Mill carried forth Milton's metaphor in his opposition to government censorship and expressly adopted a free trade in ideas conception of the free speech guarantee:

> [T]he peculiar evil of silencing the expression of an opinion is, that is robbing the human race; posterity as well as the existing generation; those who dissent

Laski, and Professor Zechariah Chafee, Jr.,[26] Holmes argued that the state should not be permitted to interfere in the free trade of ideas by prohibiting speech considered dangerous by the government, unless such interference was necessary to avert an imminent danger that was likely to occur.[27] According to this model, speech should be conceptualized in a manner analogous to other market goods, upon which market forces should be allowed to operate freely, absent an emergency. Holmes believed that the First Amendment meant that government intervention to hold in check what the government believed were dangerous ideas was unconstitutional, and that the *Abrams* defendants' criticism of government actions should have been allowed to compete with other ideas in the market and have their say, without the government intervening to censor or punish them.[28]

According to Holmes and his free speech ally Justice Louis Brandeis,[29] the proper response to allegedly harmful speech is not censorship but counterspeech. Consistent with the marketplace of ideas theory, Holmes and Brandeis believed that, absent a true emergency, the proper response to bad speech is good speech.[30] As Brandeis explained in his oft-quoted concurrence in *Whitney v. California*, joined by Holmes:

> Those who won our independence believed that [within our system of free expression] . . . discussion affords ordinarily adequate protection against the dissemination of noxious doctrine . . . . [The Founders knew 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. . . . If*

> from the opinion, still more than those who hold it. If the opinion is right, they are deprived of the opportunity of exchanging error for truth: if wrong, they lose, what is almost as great a benefit, the clearer perception and livelier impression of truth, produced by its collision with error.

JOHN STUART MILL, UTILITARIANISM, LIBERTY AND REPRESENTATIVE GOVERNMENT 79 (J.M. Dent & Sons 1922) (1861).

26 *See, e.g.*, THOMAS HEALY, THE GREAT DISSENT (2013); GEOFFREY R. STONE, PERILOUS TIMES 198–211 (2004); Gerald Gunther, Essay, *Learned Hand and the Origins of Modern First Amendment Doctrine: Some Fragments of History*, 27 STAN. L. REV. 719 (1975); David M. Rabban, *The Emergence of Modern First Amendment Doctrine*, 50 U. CHI. L. REV. 1205, 1207 (1983).

27 *See* Schenck v. United States, 249 U.S. 47, 52 (1919).

28 *Abrams*, 250 U.S. at 630 (Holmes, J., dissenting).

29 Although Holmes and Brandeis were united in their commitment to strong free speech protections, they differed in their understanding of the philosophical foundations for such protections. Holmes, as discussed above, adopted a utilitarian conception, emphasizing the importance of the free trade in ideas, while Brandeis focused in particular on the importance of free speech for the deliberative democratic process. *See, e.g.*, Vincent Blasi, *The First Amendment and the Ideal of Civic Courage: The Brandeis Opinion in* Whitney v. California, 29 WM. & MARY L. REV. 653 (1988).

30 *Abrams*, 250 U.S. at 630 (Holmes, J. dissenting). As Holmes explained in his *Abrams* dissent, "[o]nly the emergency that makes it immediately dangerous to leave the correction of evil counsels to time warrants making an exception to the sweeping command, 'Congress shall make no law . . . abridging the freedom of speech.'" *Id.* at 630–31 (omission in original).

> *there be time to expose through discussion the falsehood and fallacies [of speech], to avert the evil by the process of education, the remedy to be applied is more speech, not enforced silence.*[31]

According to the marketplace theory, ideas should be allowed to compete freely in the marketplace unfettered by government intervention, absent emergency conditions. The remedy for harmful ideas in this marketplace is not censorship but counterspeech, which works by allowing those who are exposed to bad speech to be exposed to good speech as a counterweight.

While the marketplace of ideas theory accords broad protection to good and bad *ideas*, it does not accord the same broad protections to good and bad assertions of *fact*. Although the Supreme Court in embracing the marketplace of ideas theory has made clear that there is no such thing as a false *idea*—and that all *ideas* are protected—it has also emphasized that false statements of *fact* are not similarly immune from regulation. While the Court has sometimes recognized the minimal potential contributions to the marketplace of ideas made by harmless lies[32] or false statements of fact,[33] it has also emphasized that the First Amendment does not stand in the way of regulating intentionally false, harmful assertions of fact.[34] After *New York Times Co. v. Sullivan*, false statements of fact that are made with knowledge of their falsity or with reckless disregard as to their truth or falsity, and that harm another's reputation, are actionable.[35]

Holmes and Brandeis advanced the marketplace model of free speech at a time when the marketplace for speech was radically different than today's marketplace, and when government intervention into economic markets generally was rare.[36] The prevailing marketplace for speech in their time centered around speakers on soapboxes and printed works like newspapers, journals, and leaflets like those tossed by the *Abrams* defendants from the upper floor of buildings in an attempt to widely disseminate their ideas.[37]

---

31 Whitney v. California, 274 U.S. 357, 375–77 (1927), (Brandeis, J., concurring) (emphasis added), *overruled in part by* Brandenburg v. Ohio, 395 U.S. 444, 447 (1969) (per curiam).

32 *See* United States v. Alvarez, 567 U.S. 709, 732 (2012) (Breyer, J., concurring in the judgment) (arguing that when Alvarez posed as a military medal recipient, this was a seemingly harmless lie, since this did not hurt anyone and was a lie that could be easily falsified if a list of medal recipients were made available on the internet).

33 *See* New York Times Co. v. Sullivan, 376 U.S. 254 (1964).

34 *See* Gertz v. Robert Welch, Inc., 418 U.S. 323, 340 (1974) ("[T]here is no constitutional value in false statements of fact. Neither the intentional lie nor the careless error materially advances society's interest in 'uninhibited, robust, and wide-open' debate on public issues." (quoting *Sullivan*, 376 U.S. at 270)).

35 *See Sullivan*, 376 U.S. at 279–80.

36 *See* Robert Higgs, *The Growth of Government in the United States*, FOUND. FOR ECON. EDUC. (Aug. 1, 1990), https://fee.org/articles/the-growth-of-government-in-the-united-states (contrasting the government taxes in the early twentieth century, six to seven percent of the gross national product, with 1950 taxes, twenty-four percent of the gross national product).

37 *See* Abrams v. United States, 250 U.S. 616, 617–18 (1919).

Such markets provide a sharp contrast to the online marketplaces for speech characteristic of today, in which Russian troll farms create false online accounts on Twitter and Facebook to widely disseminate false news stories at a staggering level of magnitude to sabotage the democratic process in the United States.[38] Holmes himself downplayed the potential harms inherent in the unregulated marketplace of ideas of the early twentieth century, characterizing the speech at issue in *Abrams*, for example, as "silly leaflet[s]" disseminated by "poor and puny anonymities," with little chance of influencing the populace or ultimately winning out in the marketplace.[39] In contrast, as Part II describes in greater detail, although the speech at issue today may be advanced by "anonymities," the speakers are not "puny" and the unregulated marketplace of ideas in 2019 is anything but harmless. As the harms embodied in today's marketplace of ideas become increasingly more severe, the arguments for intervention to fix the flaws in this marketplace become more compelling.

In addition, since 1919, government intervention to correct market failures in economic markets has become increasingly more frequent and accepted.[40] While government intervention in speech markets is and should continue to be subject to far more searching scrutiny than intervention in economic markets, modern First Amendment jurisprudence does not render the government powerless to provide narrowly tailored remedies directed to fixing the flaws in today's marketplace of ideas. Of course, our hands are not as free as in other countries, which, like Germany, have the power to enact regulations directed to addressing the flaws in the online marketplace of ideas unfettered by the First Amendment's constraints. Yet, consistent with the First Amendment's dictates, the U.S. government enjoys certain limited but powerful avenues for attempting to address these flaws. Below, this Article examines in greater detail the flaws in the online marketplace of ideas and evaluates the efficacy and constitutionality of tools that the government, online platforms, and affected individuals are wielding to attempt to remedy such problems.

II. The Unique Problems of Today's Online Marketplace of Ideas

Today's online marketplace of ideas is fraught with unique problems. First, as discussed above, in the online marketplace of ideas, individuals are increasingly siloed in their own echo chambers to an unprecedented degree, such that counterspeech may be of limited effect.[41] Second, the online infor-

38 *See, e.g.*, Maza, *supra* note 6.

39 *Abrams*, 250 U.S. at 628–29 (Holmes, J., dissenting).

40 *See, e.g.*, David Brodwin, *Why We Need the Government in the Marketplace*, U.S. News & World Rep. (Dec. 21, 2012), https://www.usnews.com/opinion/blogs/economic-intelligence/2012/12/21/why-we-need-the-government-in-the-marketplace.

41 *See, e.g.*, Philip M. Napoli, *What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble*, 70 Fed. Comm. L.J. 55, 77 (2018) (describing the "essence of the filter bubble phenomenon" as "the intertwining of individual and algorithmic content personalization on social media and other news aggregation platforms

mation ecosystem is besieged by false news and intentional misinformation.[42] In particular, a dire problem that today's online marketplace of ideas faces is the manipulation of speech forums like Facebook and Twitter by foreign operatives in an attempt to interfere with our democratic processes, as occurred in the context of large-scale Russian interference in the 2016 presidential election and the 2018 midterm elections in the United States.[43] Such false and intentionally misleading information on the internet is particularly problematic, given that the internet is a dominant (if not *the* dominant) source of information in the political sphere, with two-thirds of Americans identifying internet sources as their leading sources of information in connection with the most recent U.S. presidential election.[44]

The influence of misinformation and foreign interference in our 2016 presidential election was exacerbated by the use of automation in the form of bots, trolls, and fake accounts to amplify disinformation, manipulate public discourse, exacerbate political and social divisions, and deceive voters on a mass scale, especially via Twitter platforms, in a manner that was targeted to members of the U.S. electorate, especially in swing states.[45] Such automated accounts have the ability to tweet messages out hundreds of times per day and to drown out the voices of U.S. citizens.[46] Russian bots, for example, were responsible for a substantial percentage of election-related tweets

---

work[ing] to deflect news sources and content that do not correspond to the user's established content preferences and political orientation" (footnote omitted)).

42 *See id.* at 70 (comparing the cost of production between fake news and legitimate news).

43 *See* OFFICE OF THE DIR. OF NAT'L INTELLIGENCE, *supra* note 5, at 1; Rebecca Ballhaus & Dustin Volz, *U.S. Intelligence Officials Warn of 'Pervasive' Russian Efforts to Disrupt 2018 Elections*, WALL ST. J., https://www.wsj.com/articles/u-s-intelligence-officials-warn-of-pervasive-russian-efforts-to-disrupt-2018-elections-1533235652 (last updated Aug. 3, 2018); Jonathon Morgan & Ryan Fox, *Russians Meddling in the Midterms? Here's the Data*, N.Y. TIMES (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/opinion/midterm-elections-russia.html (suggesting that social media platforms' measures "may have rendered some of the Russian tactics of 2016 less effective" but also may "have merely caused Russia to shift or develop new tactics," contributing to "more overall activity in real time from continuing Russian online influence operations targeting the midterm elections than has been disclosed by social media platforms").

44 *See* Honest Ads Act, S. 1989, 115th Cong. § 3(10) (2017); PEW RESEARCH CTR., ELECTION 2016: CAMPAIGNS AS A DIRECT SOURCE OF NEWS 28 (2016).

45 Natasha Bertrand, *Twitter Users Spreading Fake News Targeted Swing States in the Run-Up to Election Day*, BUS. INSIDER (Sept. 28, 2017), https://www.businessinsider.com/fake-news-and-propaganda-targeted-swing-states-before-election-2017-9.

46 *See, e.g.*, Philip N. Howard et al., *Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration*, 15 J. INFO. TECH. & POL. 81, 83–84 (2018) ("Social bots . . . are social media accounts equipped with algorithms that post, tweet, or message of their own accord. . . . In addition, a growing amount of computationally intensive social science has demonstrated that bots can have a political impact, not so much in changing voter opinion but in attacking journalists and discrediting political leaders . . . .").

directed to the swing states of Pennsylvania, Michigan, and Wisconsin,[47] as well as to the battlegrounds states of Ohio, Missouri, Florida, North Carolina, and Colorado, during the 2016 presidential election.[48] False news available on online platforms—especially Facebook—played a significant role in influencing members of the electorate leading up to the 2016 election. More than one quarter of voting-age adults visited a false news website in the final weeks of the 2016 campaign.[49] Indeed, in the months leading up to the election, the top twenty fake news stories had more "engagements" (which includes shares, reactions, and comments) on Facebook (with 8.7 million engagements) than the twenty top hard news stories (with 7.3 million engagements).[50] In the final three months of the U.S. presidential campaign, the top performing fake election news stories on Facebook generated more engagements than the top stories from major news outlets such as the *New York Times*, the *Washington Post*, *Huffington Post*, and *NBC News*,[51] and material generated by the Kremlin reached 126 million American Facebook users.[52] The top five false news stories were all pro-Trump and anti-Clinton, as depicted in Figure 1.[53]

47 *See* Philip N. Howard et al., *Social Media, News and Political Information During the US Election: Was Polarizing Content Concentrated in Swing States?* (Project on Computational Propoganda, Working Paper No. 2017.8, 2017), https://comprop.oii.ox.ac.uk/research/working-papers/social-media-news-and-political-information-during-the-us-election-was-polarizing-content-concentrated-in-swing-states/. Trump won the Electoral College because some eighty thousand votes went his way in Wisconsin, Michigan, and Pennsylvania. *See, e.g.*, KATHLEEN HALL JAMIESON, CYBERWAR: HOW RUSSIAN HACKERS AND TROLLS HELPED ELECT A PRESIDENT 67 (2018).

48 *See* Bertrand, *supra* note 45 (noting that Russian bots targeted battleground states); Charlie Mahtesian, *What Are the Swing States in 2016?*, POLITICO (June 15, 2016), https://www.politico.com/blogs/swing-states-2016-election/2016/06/what-are-the-swing-states-in-2016-list-224327 (listing swing states for the 2016 presidential election); *see also* Scott Jaschik, *How Russian Bots Spread Fear at University in the U.S.*, INSIDE HIGHER ED (Feb. 15, 2018), https://www.insidehighered.com/news/2018/02/15/journal-article-explains-how-russian-bots-created-fear-university-missouri (explaining how Russian bots targeted Missouri).

49 *See* Danielle Kurtzleben, *Did Fake News on Facebook Help Elect Trump? Here's What We Know*, NPR (Apr. 11, 2018), https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election.

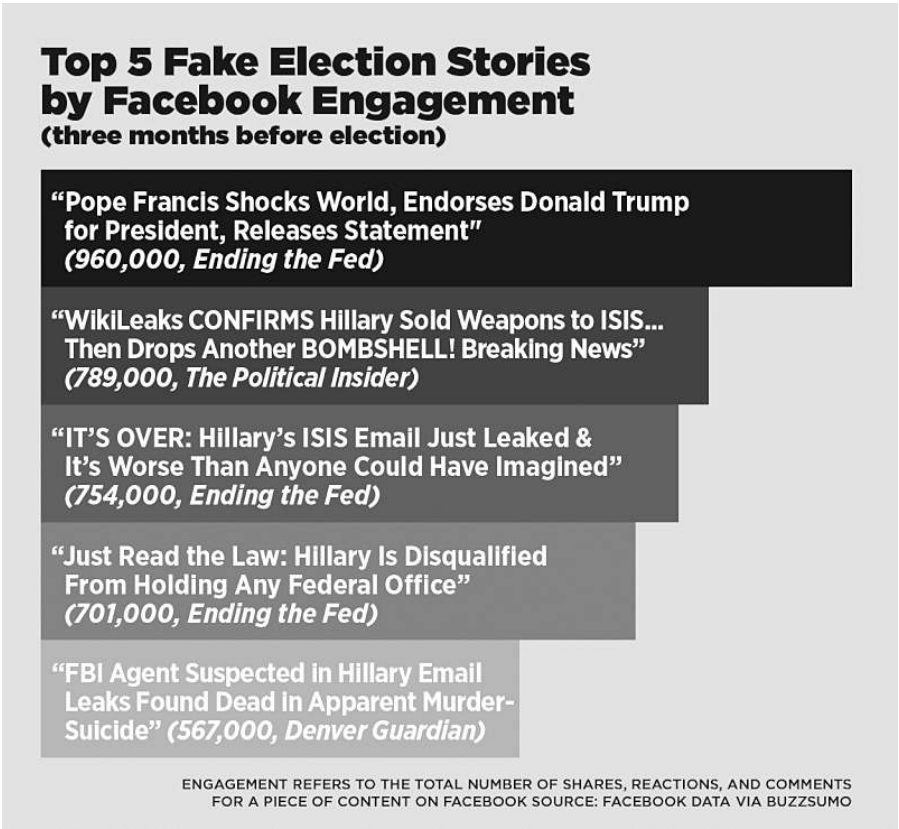
50 *Id.*

51 *See* Craig Silverman, *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook*, BUZZFEED NEWS (Nov. 16, 2016), https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook#.emA15rzd0.

52 *See* Jane Mayer, *How Russia Helped Swing the Election for Trump*, NEW YORKER (Oct. 1, 2018), https://www.newyorker.com/magazine/2018/10/01/how-russia-helped-to-swing-the-election-for-trump/.

53 *See* Silverman, *supra* note 51.

FIGURE 1



Twitter was a primary target of Russia's false news and misinformation offensives during the 2016 election, as a St. Petersburg–based troll factory known as the Internet Research Agency used Twitter as a vehicle to create fake accounts to exacerbate political and social tensions in the United States and to mislead U.S. voters.[54] The Internet Research Agency controlled more than three thousand Twitter accounts during the 2016 U.S. elections, and another fifty thousand automated accounts were connected to the Russian government,[55] creating interferences that Twitter was unprepared to handle.[56]

54 *See* Craig Timberg & Elizabeth Dwoskin, *Twitter Is Sweeping Out Fake Accounts Like Never Before, Putting User Growth at Risk*, WASH. POST (July 6, 2018), https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/?noredirect=on&utm_term=.06f5316227a8.

55 *Id.*

56 As Twitter CEO Jack Dorsey acknowledged during the September 2018 congressional hearings, "[w]e found ourselves unprepared and ill-equipped for the immensity of the problems we've acknowledged," including "[a]buse, harassment, troll armies, propaganda through bots and human coordination, disinformation campaigns and divisive filter

False news, fake sites, and amplification originated by Russian operatives continue to distort our information ecosystem, and these foreign operatives have achieved their goals so successfully that domestic operatives with similar motives are following their lead. In the context of the partisan debate surrounding the Senate confirmation of Justice Brett Kavanaugh to the Supreme Court, for example, domestic sources of misinformation adopted foreign operatives' strategies to spread lies about Kavanaugh-accuser Christine Blasey Ford and her attorneys. In conjunction with Ford's testimony before the Senate, the conservative website Right Wing News posted several false news stories about Ford on its website and then deployed various Facebook pages and accounts to proliferate these false stories.[57]

False news, misinformation, and conspiracy theories have also proliferated online outside of the election context, as in the case of Alex Jones's InfoWars conspiracy theories claiming that the Sandy Hook massacre never occurred and that the government and victims' families staged the news events surrounding the massacre in an attempt to limit Americans' Second Amendment rights.[58]

In sum, today's online marketplace of ideas is besieged by false news and intentional misinformation, creating a host of problems for our modern democracy, in which citizens increasingly rely upon the internet in general—and social media sites like Twitter and Facebook in particular—for accessing news and information.

III. Fixing the Flaws in the Online Marketplace of Ideas

A. The European Union's Approach, the German Approach, and Potential Regulatory Spillover to the United States

In response to the profound problems besieging the online marketplace of ideas discussed above, it might be tempting to wave a magic wand to make such unwanted content instantly disappear. This is essentially the approach

bubbles." Tony Romm & Craig Timberg, *Facebook and Twitter Testified Before Congress. Conservative Conspiracy Theorists Lurked Behind Them.*, Wash. Post (Sept. 5, 2018), https://www.washingtonpost.com/technology/2018/09/05/facebook-twitter-sandberg-dorsey-congress-tech-hearings/?utm_term=.862c1ab2b3a5.

57 As the *New York Times* reports, while "domestic sites are emulating the Russian strategy of 2016 by aggressively creating networks of Facebook pages and accounts—many of them fake—that make it appear as if the ideas they are promoting enjoy widespread popularity." Sheera Frenkel, *Facebook Tackles Rising Threat: Americans Aping Russian Schemes to Deceive*, N.Y. Times (Oct. 11, 2018), https://www.nytimes.com/2018/10/11/technology/fake-news-online-disinformation.html?module=inline.

58 *See* Chris Sommerfeldt, *See It: Right-Wing Conspiracy Theorist Alex Jones Doubles Down on 'Completely Fake' Sandy Hook Massacre Claims*, N.Y. Daily News (Nov. 18, 2016), http://www.nydailynews.com/news/politics/alex-jones-doubles-completely-fake-sandy-hook-claims-article-1.2878305#; *see also* Vanessa Romo, *Sandy Hook Parents Sue Conspiracy Theorist Alex Jones over Claim Shooting Was 'Fake,'* NPR (Apr. 17, 2018), https://www.npr.org/sections/thetwo-way/2018/04/17/603223968/sandy-hook-parents-sue-conspiracy-theorist-alex-jones-over-claim-shooting-was-fa.

recently adopted by Germany, which builds on the European Union's approach and which starts from a different calculus of the harms caused by free speech as weighed against its benefits.[59] The European approach to fixing the flaws in the online marketplace of ideas, which is dramatically different from the United States' approach, imposes substantial pressure on social media platforms to monitor their content. Consequently, platforms may end up responding to these pressures in ways that affect what U.S. audiences can view online.

The current German legislative approach to fixing the flaws in the online marketplace of ideas builds upon an agreement—the EU Code of Conduct—that Facebook, Twitter, YouTube, and Microsoft entered into with the European Commission in 2016 (and that Google and Instagram have recently joined) in which these companies agreed to rapidly remove online hate speech upon receiving notice of the same.[60] (In the European Union—unlike in the United States under the Communications Decency Act (CDA) 230 regime[61]—online platforms can be held legally responsible for the illegal content they host if they do not "expeditiously . . . remove or . . . disable access" to such content.)[62] Under the EU Code of Conduct, the companies agree to remove from their platforms, within twenty-four hours of notification, "illegal hate speech"—defined as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin."[63] In a similar vein, the European Union has also recently secured a commitment from Google, Facebook, and other online platforms to adopt self-regulatory measures to address the spread of fake news and disinformation.[64] German lawmakers claimed that social media companies

59 For other examples of extreme approaches to the spread of fake news online, consider Saudi Arabia's approach, under which the sharing or spreading of fake news or rumors that might affect public order and security is punishable by a five-year prison term and three million Saudi riyal fine. *See 5-Year Jail, 3 Million Fine for Rumormongers*, SAUDI GAZETTE (Oct. 13, 2018), http://saudigazette.com.sa/article/545523.

60 *See* European Commission Press Release IP/18/261, Countering Illegal Hate Speech Online—Commission Initiative Shows Continued Improvement, Further Platforms Join (Jan. 19, 2018); European Commission Press Release IP/16/1937, European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech (May 31, 2016).

61 47 U.S.C. § 230 (2012); *see also Section 230 of the Communications Decency Act*, ELECTRONIC FRONTIER FOUND., https://www.eff.org/issues/cda230 (last visited Dec. 14, 2018) (summarizing the impact of § 230).

62 *See* Council Directive 2000/31, pmbl. ¶ 46, 2000 O.J. (L 178) 6 (EC).

63 *See* EUROPEAN COMM'N, CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE 1 (2016), https://ec.europa.eu/info/sites/info/files/code_of_conduct_on_countering_illegal_hate_speech_online_en.pdf.

64 In April, the European Commission instructed tech firms, including Facebook and Google, to draft a code of practice to combat misleading and illegal content or face further regulation. In September 2018, the platforms agreed to a voluntary code, under which they would reject payment from sites that spread fake news and distinguish advertisements from editorial content, among other measures. *See* Foo Yun Chee, *Facebook, Google to Tackle*

were not acting quickly and effectively enough to comply with the 2016 EU Code of Conduct and argued that national legislation was necessary to hold social media companies to account. In response, in 2017 Germany enacted legislation requiring social media companies to take swift, drastic, and censorial actions in response to false news, hate speech, and similar harmful information online that arguably distorts the online marketplace of ideas.[65] Other European countries—and the European Union as a body itself—are considering following suit.[66] Germany's parliament approved this legislation—the Network Enforcement Act (in German, Netzwerkdurchsetzungsgesetz or NetzDG for short)—on June 30, 2017, and the law took effect, after a transitional period, on January 1, 2018.[67] The law requires that social media companies do the following: (1) block access within Germany to "manifestly unlawful" content within twenty-four hours of receiving notice of any such content, and (2) block access to other unlawful content (that is unlawful but not "manifestly" unlawful) within seven days of receiving such notice.[68] The NetzDG legislation imposes fines of up to fifty million euros for noncompliance.[69] Unlawful categories of speech that must be promptly blocked includes content that is unlawful under various sections of the German Criminal Code, including laws prohibiting hate speech, "public incitement to crime," "dissemination of depictions of violence," forming criminal or terrorist organizations, and defamation,[70] to name a few. The law does

Spread of Fake News, Advisors Want More, REUTERS (Sept. 26, 2018), https://www.reuters.com/article/us-eu-tech-fakenews/facebook-google-agree-to-tackle-fake-news-eu-idUSKCN1M61AG; *Code of Practice on Disinformation*, EUROPEAN COMM'N (Sept. 26, 2018), https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

65 *See Overview of the NetzDG Network Enforcement Law*, *supra* note 11.

66 Emma Thomasson, *Germany Looks to Revise Social Media Law as Europe Watches*, REUTERS (Mar. 8, 2018), https://www.reuters.com/article/us-germany-hatespeech/germany-looks-to-revise-social-media-law-as-europe-watches-idUSKCN1GK1BN; *see also Germany Starts Enforcing Hate Speech Law*, BBC (Jan. 1, 2018), https://www.bbc.com/news/technology-42510868 ("The German law is the most extreme example of efforts by governments and regulators to rein in social media firms.").

67 *See Flawed Social Media Law*, *supra* note 12.

68 *See id.*

69 *See id.*

70 Sections of the German Criminal Code referenced by the NetzDG legislation include: Dissemination of propaganda material of unconstitutional organizations (section 86), Using symbols of unconstitutional organizations (section 86a), Preparation of a serious violent offence endangering the state (section 89a), Defamation of the president (section 90), Defamation of the state and its symbols (section 90a), Anti-constitutional defamation of constitutional organs (section 90b), Encouraging the commission of a serious violent offence endangering the state (section 91), Treasonous forgery (section 100a), Public incitement to crime (section 111), Breach of the public peace by threatening to commit offences (section126), Forming criminal and terrorist organizations, domestically and abroad (sections 129–129b) Incitement to hatred (section 130), Dissemination of depictions of violence (section 131), Rewarding and approving of offences (section 140), Defamation of religions, religious and ideological associations (section 166), Distribution, acquisition, and possession of child pornography (section 184b), Distribution of pornographic performances by broadcasting, media services, or telecommunications services

not provide for any remedy for those whose content is wrongfully blocked pursuant to NetzDG's mandates, other than to complain to the social media provider who blocked their content.[71]

Facebook, Twitter, and other large social media companies have been scrambling to comply with the dictates of NetzDG since it became fully effective in January 2018.[72] To implement the legislation's mandates, Facebook adopted new reporting mechanisms and hired thousands of employees to review users' reports of unlawful content.[73] It also introduced new features to flag controversial content and has spent months training its employees on NetzDG's requirements.[74]

Not surprisingly, NetzDG has been the subject of intense debate in Germany in the months since its passage and has been criticized for imposing a censorship regime, as well as for backfiring and actually increasing support for groups and causes that have been censored under this regime.[75] Facebook has implemented the legislation's mandate by blocking thousands of posts, including posts made by public officials and posts on matters that are arguably of legitimate interest to the public. For example, among the very first posts that Facebook blocked once the law became effective was a comment made by a government official criticizing the actions of the Cologne police for posting a New Year's greeting that provided safety advice in German, English, French, and Arabic.[76] In a New Year's Eve post by the deputy leader of Germany's far-right Alternative für Deutschland (AfD) party, Beatrix von Storch asked, "What the hell is happening in this country? Why is an official police site tweeting in Arabic? Do you think it is to appease

(section 184d), Insult and defamation (sections 185–187), Causing the danger of criminal prosecution by informing on a person (section 241), and Forging of data intended to provide proof (section 269). *See* Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], June 30, 2017, BUNDESGESETZBLATT [BGBI].

71 *See* Thomasson, *supra* note 66.

72 *See* Philip Oltermann, *Tough New German Law Puts Tech Firms and Free Speech in Spotlight*, GUARDIAN (Jan. 5, 2018), https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight.

73 *See Flawed Social Media Law*, *supra* note 12; *see also Network Enforcement Act ("NetzDG")*, FACEBOOK HELP CTR., https://www.facebook.com/help/285230728652028 (last visited Sept. 28, 2018).

74 *See* Oltermann, *supra* note 72; *see also What Happens After I Submit a Report Under the NetzDG?*, FACEBOOK HELP CTR., https://www.facebook.com/help/499717357069620 (last visited Sept. 28, 2018) (explaining to users what happens after they report alleged NetzDG violations).

75 *See* Linda Kinstler, *Germany's Attempt to Fix Facebook Is Backfiring*, ATLANTIC (May 18, 2018), https://www.theatlantic.com/international/archive/2018/05/germany-facebook-afd/560435 ("Germany's attempt to regulate speech online has seemingly amplified the voices it was trying to diminish.").

76 *See id.*; *see also* Philip Oltermann & Pádraig Collins, *Two Members of Germany's Far-Right Party Investigated by State Prosecutor*, GUARDIAN (Jan. 2, 2018), https://www.theguardian.com/world/2018/jan/02/german-far-right-mp-investigated-anti-muslim-social-media-posts.

the barbaric, gang-raping hordes of Muslim men?"[77] Referencing the German penal code's prohibition on speech that constitutes "incitement of the people," Facebook blocked the deputy leader's post.[78] Facebook has also blocked posts that are critical of the AfD party and its leadership, including a post that referred to the coleader of the party as a "Nazi pig."[79] Critics have argued that NetzDG implements "Stasi methods" that are "reminiscent of censorship in communist East Germany."[80] Opponents of the legislation also contend that Facebook's "delete when in doubt" practice has a chilling effect on speech both online and offline,[81] and that in the NetzDG era, "people are more careful what to think, what to write" and "[l]ots of people are afraid of losing of their accounts."[82] Critics also complain about the process through which such censorship occurs. They lament the fact that NetzDG has created a regime in which "companies can play judges"[83] and through which the legislation "outsourc[es] censorship to private companies and infring[es] on civil liberties."[84] As one public official complained, under NetzDG, "too many competences that require legal expertise are delegated to tech companies."[85] In addition, critics claim that NetzDG has been enforced arbitrarily and in a politically biased manner,[86] and that the law has had the counterproductive effect of shoring up support for those who have been censored. In particular, opponents of the law lament the fact that Alternative für Deutschland is "using the law to paint itself as a victim" and that its leaders are now characterizing themselves as "opinion martyrs."[87]

Facebook's implementation of NetzDG's mandate has also been subject to legal challenges in which plaintiffs contend that blocking illegal content

77 *See* Chase Winter, *AfD Politician 'Censored' Under New German Hate Speech Law for Anti-Muslim Tweet*, DW NEWS (Feb. 1, 2018), https://www.dw.com/en/afd-politician-censored-under-new-german-hate-speech-law-for-anti-muslim-tweet/a-41992679.

78 *See* Oltermann, *supra* note 72.

79 *See* Jens Meyer-Wellmann, *Alice Weidel Could Bring Facebook to Its Knees*, HAMBURGER ABENDBLATT (Apr. 27, 2018), https://www.abendblatt.de/hamburg/article214139039/Facebook-droht-Niederlage-gegen-AfD-Politikerin-Alice-Weidel.html.

80 *See* Oltermann, *supra* note 72.

81 *See* Sven Jacobs, *Already Changes to the New German Law on Hate Speech on Social Media on the Horizon?*, SOC. MEDIA L. BULL. (Mar. 20, 2018), https://www.socialmedialawbulletin.com/2018/03/already-changes-new-german-law-hate-speech-social-media-horizon; Kinstler, *supra* note 75.

82 Kinstler, *supra* note 75 (quoting German internet lawyer Jeorg Heidrich).

83 Mark Scott & Janosch Delcker, *Free Speech vs. Censorship in Germany*, POLITICO (Jan. 4, 2018), https://www.politico.eu/article/germany-hate-speech-netzdg-facebook-youtube-google-twitter-free-speech.

84 Christof Kerkmann, *German Court Overturns Facebook 'Censorship,'* HANDELSBLATT TODAY (Apr. 13, 2018), https://global.handelsblatt.com/politics/german-court-facebook-censorship-910635.

85 Oltermann, *supra* note 72 (quoting Konstantin von Notz, member of Parliament).

86 *See* Kerkmann, *supra* note 84.

87 Oltermann, *supra* note 72; *see also* Kinstler, *supra* note 75 (describing the NetzDG's attempt to regulate online speech as "amplif[ying] the voices it was trying to diminish," specifically with regard to the AfD).

within Germany is insufficient and that Facebook must block such content in such a manner that *Germans located outside of Germany* are also prohibited from accessing it.[88] Depending on the result of these challenges, Facebook may ultimately determine that the easiest way to comply with the law's mandate is to remove illegal content entirely, instead of blocking access to it within Germany *and* for Germans outside Germany.

Twitter's implementation of NetzDG has been similar to that of Facebook and has resulted in similar controversies. Like Facebook, Twitter has created mechanisms for users to flag content as potentially illegal under German law, hired new moderators to monitor these reports, and set up a legal compliance office specifically for NetzDG.[89] From January through June 2018, Twitter users flagged over two hundred fifty thousand items as illegal, and Twitter responded by removing ten percent of those items (and removed the vast majority of such items within twenty-four hours, as mandated by NetzDG).[90] The content that users flagged as illegal on Twitter and Facebook is largely similar and has sparked similar concerns that the platforms will resort to overreaction and overblocking.[91] For example, like Facebook, Twitter also removed the New Year's Eve post by Alternative für Deutschland deputy leader Beatrix von Storch,[92] and also suspended Storch's account for breaching Twitter's code of conduct.[93] Both Facebook's

88 In connection with a German *Huffington Post* article about Alice Weidel, coleader of the AfD party, a user commented that Weidel was a "Nazi pig" and made disparaging remarks about Weidel's sexual orientation. Meyer-Wellmann, *supra* note 79. Weidel demanded that the disparaging comments be blocked by Facebook, and Facebook complied by making the user's comments inaccessible to all Facebook users using a German IP address. *See* Kinstler, *supra* note 75.

89 *See* Shashi Jayakumar, *Germany's NetzDG: Template for Dealing with Fake News?*, S. RAJARATNAM SCH. OF INT'L STUD. (Mar. 13, 2018), https://www.rsis.edu.sg/rsis-publication/cens/co18041-germanys-netzdg-template-for-dealing-with-fake-news/#.W7FI52aZM6i (noting that Twitter has set up a legal compliance office specifically to handle issues arising from NetzDG); Oltermann, *supra* note 72 (noting that Twitter has been hiring "more German-language moderators with a background in law"); *see also* Thomasson, *supra* note 66 (reporting that Twitter "declined to comment on how it is implementing the law").

90 Kirsten Gollatz et al., *Removals of Online Hate Speech in Numbers*, ALEXANDER VON HUMBOLDT INSTITUT FÜR INTERNET UND GESELLSCHAFT (Aug. 9, 2018), https://www.hiig.de/en/removals-of-online-hate-speech-numbers.

91 James McAuley, *France Weighs a Law to Rein in 'Fake News,' Raising Fears for Freedom of Speech*, WASH. POST (Jan. 10, 2018), https://www.washingtonpost.com/world/europe/france-weighs-a-law-to-rein-in-fake-news-raising-fears-for-freedom-of-speech/2018/01/10/78256962-f558-11e7-9af7-a50bc3300042_story.html?utm_term=.c77faf5ba81a (reporting that Twitter removed a tweet "criticiz[ing] German authorities for allegedly failing to pursue xenophobia investigations"); *see also* Kinstler, *supra* note 75 ("[M]any other Facebook and Twitter users have seen their posts flagged or removed because of the new regulation without quite knowing why.").

92 *See* Oltermann, *supra* note 72.

93 *Id.* (highlighting the suspension of a German satirical magazine's Twitter account after the magazine, called *Titanic*, tweeted a parody of Storch's tweet); *see also* Scott & Delcker, *supra* note 83 (reporting Twitter's refusal to comment on the *Titanic* case, but quoting Twitter as stating that "under certain circumstances, potentially harmful tweets

and Twitter's actions in response to Storch's and similar posts have been subject to criticisms that the companies are engaging in censorship of political speech.[94]

Separate and apart from the debate in Germany over NetzDG and its implementation, such an approach to fixing the flaws in the online marketplace of ideas could never pass constitutional muster in the United States. Given the First Amendment's heavy presumption against prior restraints on speech as well as the First Amendment's underlying broader protections for hate speech, incendiary speech, etc., an approach like that taken in Germany—in which the government mandates that social media companies immediately block access to content in response to users' flagging such content as illegal—would stand no chance of passing constitutional muster in the United States. Notwithstanding the fact that NetzDG (and the EU Code of Conduct) require social media companies to block content *after* it has been published on the internet, these regulations nevertheless require that such content be blocked *prior* to a judicial determination of the content's illegality.[95] Under First Amendment jurisprudence, any state-mandated censorship that occurs outside the context of a judicial determination of the content's illegality—even censorship that occurs after the content is initially published—is an unconstitutional prior restraint on speech.[96] Accordingly,

---

may run against the company's existing terms and conditions and not be linked to the new German hate speech rules").

94 *See, e.g.*, Scott & Delcker, *supra* note 83.

95 *See* Kerkmann, *supra* note 84 (criticizing the NetzDG law for "outsourcing censorship to private companies and infringing on civil liberties").

96 As I explain elsewhere, regardless of whether censorship occurs ex ante—before publication or posting on the internet—or midstream—at some point after the content has been posted—extrajudicial censorship constitutes a presumptively illegal prior restraint. *See* Dawn C. Nunziato, *How (Not) to Censor: Procedural First Amendment Values and Internet Censorship Worldwide*, 42 GEO. J. INT'L L. 1123, 1143 (2011). Examples of the former include those imposed by prescreening film boards and internet filtering schemes that are imposed ex ante. *See id.* at 1144. Midstream prior restraints include those imposed after initial publication but before a judicial determination of the content's illegality. *See id.* Because midstream prior restraints are imposed in the absence of the procedural safeguards that attend a judicial determination, they are as constitutionally suspect as ex ante prior restraints. *See id.* Midstream prior restraints include state-mandated notice and take down systems like those imposed under the NetzDG legislation. *See id.* The Supreme Court has struck down midstream prior restraints in cases such as *Bantam Books, Inc. v. Sullivan*, 372 U.S. 58, 64 (1963). In that case, the Rhode Island Commission to Encourage Morality in Youth was charged by the state with investigating and recommending prosecution of booksellers for distributing books that were obscene or indecent. *Id.* at 59–60. The Commission reviewed books after they were already in circulation and notified distributors when they distributed a book that the Commission deemed objectionable. *Id.* at 61. Upon receipt of such notices, many distributors stopped further circulation of the identified works. *Id.* at 63. In reviewing the constitutionality of the Rhode Island scheme, the Supreme Court held that even though the restrictions on publication were imposed midstream—after initial circulation and distribution—the actions nonetheless effectuated an unconstitutional prior restraint. *Id.* at 64, 70. The Court explained that "[t]he separation of legitimate from illegitimate speech calls for . . . sensitive tools," *id.* (omission in original)

if the U.S. government imposed a similar mandate on digital platforms, such a mandate would be clearly unconstitutional.

Yet, given the pressures that the 2018 NetzDG law and the 2016 EU Code of Conduct are placing on social media companies to block content that is illegal in Germany and Europe in a manner that renders such content inaccessible for Germans and Europeans wherever they may be located, these international approaches to fixing the flaws in the online marketplace of ideas may well have an effect on the marketplace of ideas as it is accessed by U.S. citizens as well.

B. What Facebook Is Doing

In addition to complying with the mandates from the European Union under the EU Code of Conduct and the new EU Code of Practice on Disinformation, and from Germany under its NetzDG legislation described above, and in addition to removing content that violates its own community standards,[97] Facebook is taking a number of steps to attempt to remedy the flaws in the online marketplace of ideas.[98] As discussed above, Facebook's platform was subject to manipulation by Russian hackers who disseminated false news and disinformation on a massive scale during the period leading up to the 2016 U.S. presidential election.[99] After being subject to intense scrutiny from Congress and intense criticism from the court of public opinion in the United States, Facebook recently announced and implemented several measures to attempt to address these problems on its platform,[100]

(quoting Speiser v. Randall, 357 U.S. 513, 525 (1958)), and reiterated its "insistence that regulations of [illegal content] scrupulously embody the most rigorous procedural safeguards." *Id.* at 66. The Court observed that, under the Rhode Island scheme, "[t]he publisher or distributor is not even entitled to notice and hearing before his publications are listed by the Commission as objectionable" and that there was "no provision whatever for judicial superintendence before notices issue or even for judicial review of the Commission's determinations of objectionableness." *Id.* at 71. Accordingly, in the context of this midstream prior restraint, the "procedures of the Commission [were] radically deficient" and unconstitutional. *Id.*

97 Facebook's community standards prohibit posting the following types of content: violence and criminal behavior, suicide and self-injury, child nudity and sexual exploitation of children, sexual exploitation of adults, bullying, harassment, privacy violations, objectionable content, hate speech, violence and graphic content, adult nudity and sexual activity, sexual solicitation, cruel and insensitive content, spam, misrepresentation, and content violating others' intellectual property. *See Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/ (last visited Mar. 13, 2019). A full list of Facebook's community standards, as well as definitions of prohibited categories of content, can be found in Appendix A.

98 *See* Tessa Lyons, *Hard Questions: What's Facebook's Strategy for Stopping False News?*, FACEBOOK NEWSROOM (May 23, 2018), https://newsroom.fb.com/news/2018/05/hard-questions-false-news.

99 *See* Mayer, *supra* note 52.

100 False news or fake news is not per se in violation of Facebook's community standards. As Facebook explains, "There is . . . a fine line between false news and satire or opinion. For these reasons, we don't remove false news from Facebook but instead, signifi-

including by securing independent evaluation of certain content for its accuracy, deprioritizing inaccurate content, and penalizing purveyors of false content; securing and providing truthful counterspeech in response to false content; adopting a political advertising policy; removing content that incites imminent violence; and engaging in ongoing educational efforts. More specifically, as this Section discusses in detail below, Facebook's recently adopted measures to combat false news on its platform include partnering with independent third-party fact-checkers to evaluate posts; providing counterspeech in the form of "Related Articles"/"Additional Reporting on This" on topics similar to false articles posted on Facebook; limiting the distribution of posts from content providers who repeatedly share false news and eliminating their ability to profit; removing false posts that are intended to encourage violence; eliminating fake accounts; modifying its "news feed" functionality by prioritizing interactive content and allowing users to customize their feeds; and engaging in educational efforts regarding news literacy.[101] Section III.B discusses each of these efforts in turn below and analyzes their likely effectiveness in remedying the flaws in the online marketplace of ideas. This Section also assesses the extent to which these efforts are consistent with the marketplace of ideas theory of the First Amendment.

1. Partnering with Third-Party Fact-Checkers to Evaluate Potentially False Posts

Facebook is continuing to expand the partnership that it began in December 2016 with fact-checkers to evaluate content posted on its platform.[102] Through its fact-checking initiatives, Facebook is working with independent third-party fact-checkers, which are certified through the nonpartisan International Fact-Checking Network.[103] In the United States, the certified fact-checking organizations with whom Facebook works are the Associated Press, FactCheck.org, Lead Stories, and *PolitiFact*[104] (Facebook had added the *Weekly Standard* to these ranks for a period of time in an attempt to respond to critics who claimed that its fact-checking program was

---

cantly reduce its distribution by showing it lower in the News Feed." *False News*, FACEBOOK, https://www.facebook.com/communitystandards/false_news (last visited Dec. 3, 2018).

101 *See* Hunt Allcott et al., *Trends in the Diffusion of Misinformation on Social Media* app. at 4 tbl.1 (Nat'l Bureau of Econ. Research, Working Paper No. 25500, 2019), http://web.stanford.edu/~gentzkow/research/fake-news-trends-appx.pdf (listing all of Facebook's efforts to combat false news).

102 *See* Lyons, *supra* note 98.

103 *See id.*

104 *See* Mike Ananny, *Checking In with The Facebook Fact-Checking Partnership*, COLUM. JOURNALISM REV. (Apr. 4, 2018), https://www.cjr.org/tow_center/facebook-fact-checking-partnerships.php; *see also How Are Third-Party Fact-Checkers Selected?*, FACEBOOK HELP CTR., https://www.facebook.com/help/1599660546745980?helpref=faq_content (last visited Sept. 29, 2018); *Third-Party Fact-Checking on Facebook*, FACEBOOK HELP CTR., https://www.facebook.com/help/publisher/182222309230722 (last visited Mar. 14, 2019).

politically biased, but this publication is now defunct).[105] Facebook has expanded its fact-checking initiative to include the fact-checking of all public, newsworthy Facebook posts, including links, articles, photos, and videos.[106] The fact-checking process can be initiated by Facebook users flagging a post as being potentially false. To do so, a user clicks ". . ." next to the post he or she wishes to flag as false, then clicks "Report post," then clicks "It's a false news story," then clicks "Mark this post as false news."[107] (Alternatively, a user can click ". . ." next to a post, then click "Give Feedback on This Post," then click "False News.")[108] Once a post is flagged by a user as a potential false news story, it is submitted for evaluation to a third-party independent fact-checker.[109] While the process of evaluating posts in the past was triggered only by user flagging, Facebook now incorporates other ways of triggering such evaluation, including by providing its independent fact-checkers with the authority to proactively identify posts to review[110] as well as by using machine learning to identify potentially false posts.[111] For each piece of content up for review, a fact-checker has the option of providing one of eight different ratings: false, mixture, false headline, true, not eligible (if, for example, the post is not verifiable, opinion, etc.), satire, opinion, or prank generator.[112]

2. Related Articles/Additional Reporting as Counterspeech and Other Remedies in Response to False News

Once a third-party fact-checker has determined that a post is false, Facebook then initiates several steps. First, Facebook deprioritizes false posts in users' news feeds, i.e., the constantly updating list of stories in the middle of a user's home page (including status updates, photos, videos, links, app

105 *See* Matthew Ingram, The Weekly Standard *and the Flaws in Facebook's Fact-Checking Program*, COLUM. JOURNALISM REV. (Sept. 18, 2018), https://www.cjr.org/the_new_gatekeepers/the-weekly-standard-facebook.php.

106 *See* Antonia Woodford, *Expanding Fact-Checking to Photos and Videos*, FACEBOOK NEWSROOM (Sept. 13, 2018), https://newsroom.fb.com/news/2018/09/expanding-fact-checking.

107 *See How Do I Mark a Post as False News?*, FACEBOOK HELP CTR., https://www.facebook.com/help/572838089565953?helpref=faq_content (last visited Sept. 29, 2018).

108 *See* FACEBOOK, https://www.facebook.com (last visited Mar. 14, 2019).

109 *See* Lyons, *supra* note 14 ("[W]hen people on Facebook submit feedback about a story being false or comment on an article expressing disbelief, these are signals that a story should be reviewed.").

110 *See id.* ("Independent third-party fact-checkers review the stories, rate their accuracy, and write an article explaining the facts behind their rating.").

111 *See* Dan Zigmond, *Machine Learning, Fact-Checkers and the Fight Against False News*, FACEBOOK NEWSROOM (Apr. 8, 2018), https://newsroom.fb.com/news/2018/04/inside-feed-misinformation-zigmond/?url=https://newsroom.fb.com/news/2018/04/inside-feed-misinformation-zigmond/&utm_source=gg&utm_medium=ps&utm_campaign=1408779249&utm_content=279529609338&utm_term=%2Bfacebook%20%2Bfact%20%2Bchecking.

112 *Third-Party Fact-Checking on Facebook*, *supra* note 104.

activity, and likes)—such that future views of each false post will be reduced by an average of eighty percent.[113] Second, Facebook commissions a fact-checker to write a related article or "Additional Reporting on This," setting forth truthful information about the subject of the false post and the reasons why the fact-checker rated the post as false.[114] Such content is then displayed in conjunction with the false post on the same subject.[115] While Facebook formerly flagged false news sites with a "Disputed" flag, the company changed its approach in response to research suggesting that such flags may actually entrench beliefs in the disputed posts.[116] Facebook now provides "Related Articles"/"Additional Reporting on This" in conjunction with false news stories (which apparently does not result in similar entrenchment).[117] In addition, users who attempt to share the false post will be notified that the post has been disputed and will be informed of the availability of a "Related Article"/"Additional Reporting on This," as will users who earlier shared the false post,[118] as in Figure 2 below.[119]

113 *See id.*; *see also* Tessa Lyons, *Increasing Our Efforts to Fight False News*, FACEBOOK NEWSROOM (June 21, 2018), https://newsroom.fb.com/news/2018/06/increasing-our-efforts-to-fight-false-news/.

114 Geoffrey A. Fowler, *I Fell for Facebook Fake News. Here's Why Millions of You Did, Too.*, WASH. POST. (Oct. 18, 2018), https://www.washingtonpost.com/technology/2018/10/18/i-fell-facebook-fake-news-heres-why-millions-you-did-too/?noredirect=on&utm_term=.8b17ea23b1c2; *see* Tessa Lyons, *Replacing Disputed Flags with Related Articles*, FACEBOOK NEWSROOM (Dec. 20, 2017), https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation.

115 *See* Lyons, *supra* note 114; *see also* Fowler, *supra* note 114 (describing steps undertaken by Facebook to respond to fake video, including posting "Additional Reporting on This," with links to reports from fact-checking organizations); Lyons, *supra* note 14; Sara Su, *New Test with Related Articles*, FACEBOOK NEWSROOM (Apr. 25, 2017), https://newsroom.fb.com/news/2017/04/news-feed-fyi-new-test-with-related-articles.

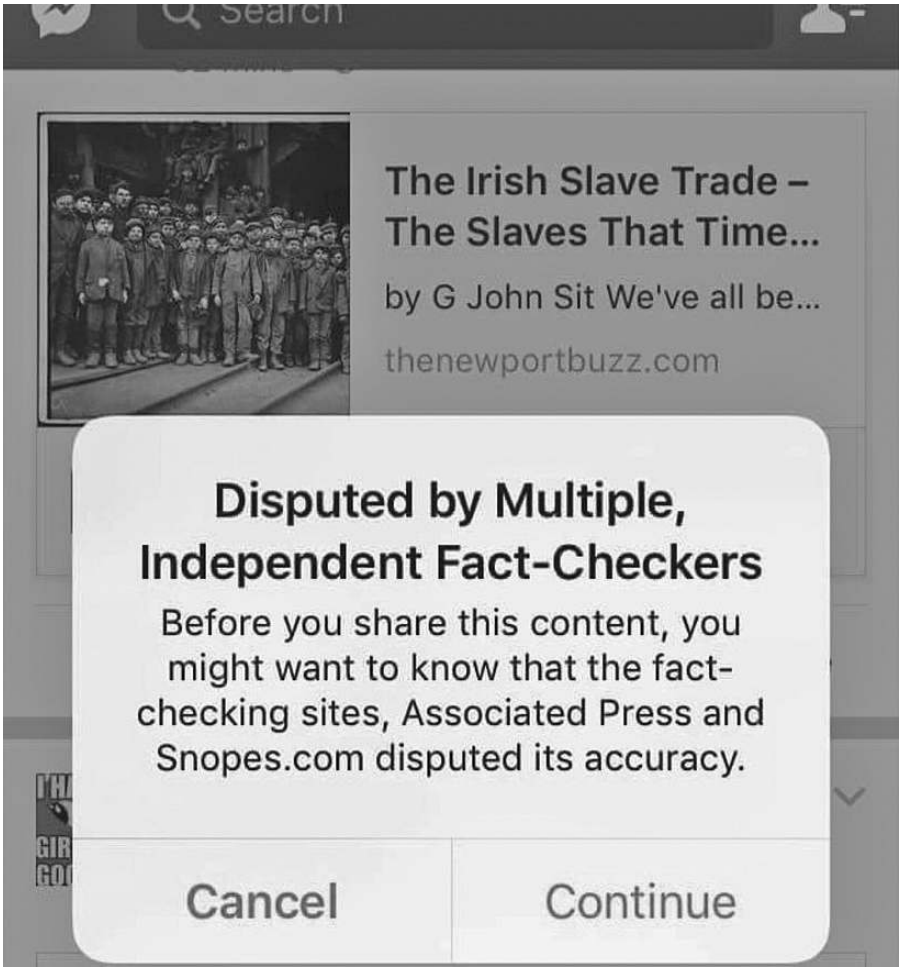
116 *See* Catherine Shu, *Facebook Will Ditch Disputed Flags on Fake News and Display Links to Trustworthy Articles Instead*, TECHCRUNCH (Dec. 20, 2017), https://techcrunch.com/2017/12/20/facebook-will-ditch-disputed-flags-on-fake-news-and-display-links-to-trustworthy-articles-instead/.

117 *See* Lyons, *supra* note 114 ("Academic research on correcting misinformation has shown that putting a strong image, like a red flag, next to an article may actually entrench deeply held beliefs . . . . Related Articles, by contrast, are simply designed to give more context, which our research has shown is a more effective way to help people get to the facts. . . . [W]e've found that when we show Related Articles next to a false news story, it leads to fewer shares than when the Disputed Flag is shown.").

118 *See* Lyons, *supra* note 14.

119 Elle Hunt, *'Disputed by Multiple Fact-Checkers': Facebook Rolls Out New Alert to Combat Fake News*, GUARDIAN (Mar. 21, 2017), https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news.

FIGURE 2



Third, content providers—i.e., Facebook pages and domains—that repeatedly publish and/or share false posts will lose their ability to monetize and advertise unless and until they issue corrections or successfully dispute fact-checkers' determinations that their posts are false.[120]

120 *See* Satwik Shukla & Tessa Lyons, *Blocking Ads from Pages That Repeatedly Share False News*, FACEBOOK NEWSROOM (Aug. 28, 2017), https://newsroom.fb.com/news/2017/08/blocking-ads-from-pages-that-repeatedly-share-false-news.

3. Transparency and Disclosure Requirements Regarding Political/ Electioneering Advertisements

Facebook also implemented a political advertising policy in May 2018.[121] The political advertising policy requires, first, that every election-related and issue advertisement made available on Facebook to users in the United States be clearly labeled as a "Political Ad" and include a "paid for by" disclosure, with the name of the individual or organization who paid for the advertisement at the top of the advertisement.[122] Second, under the policy, Facebook will collect and maintain a publicly available archive of political advertisements, which will provide information regarding the campaign budget associated with each individual ad and how many people saw it, including their ages, locations, and genders.[123] Third, under the policy, Facebook will prohibit foreign entities from purchasing political ads directed at U.S. audiences.[124] Facebook will implement this prohibition by mailing prospective political advertisers a postcard sent to a U.S. address in order to verify U.S. residency. If a prospective purchaser of a political ad is not verified under this process, it will not be able to post a political advertisement on Facebook and will be blocked from purchasing political ads in the future. Commenting on the recently implemented political advertising policy, Facebook's CEO, Mark Zuckerberg, explained, "These changes won't fix everything, but they will make it a lot harder for anyone to do what the Russians did during the 2016 election and use fake accounts and pages to run ads."[125]

Facebook's recently implemented measures imposing disclosure requirements on political ads and limiting foreign entities from purchasing political ads go beyond those that are encompassed in the proposed Honest Ads Act, discussed below,[126] and manifest a commitment from Facebook to take seriously its responsibility to address the problems of foreign interference in the U.S. political process.

4. Removing False Posts Intended and Likely to Encourage Violence

In addition to demoting posts that have been determined by fact-checkers to be false, for some types of particularly harmful content, Facebook will remove content altogether. In response to criticism that it contributed to violence against religious and ethnic minorities in Sri Lanka, Myanmar, and

121 *See* Rob Goldman & Alex Himel, *Making Ads and Pages More Transparent*, FACEBOOK NEWSROOM (Apr. 6, 2018), https://newsroom.fb.com/news/2018/04/transparent-ads-and-pages/.

122 *See id.*

123 *See* Rob Leathern, *Shining a Light on Ads with Political Content*, FACEBOOK NEWSROOM (May 24, 2018), https://newsroom.fb.com/news/2018/05/ads-with-political-content/.

124 *See Ads Related to Politics or Issues of National Importance*, FACEBOOK BUS., https://www.facebook.com/business/help/167836590566506 (last visited Mar. 14, 2019).

125 Josh Constine, *Facebook and Instagram Launch US Political Ad Labeling and Archive*, TECHCRUNCH (May 24, 2018), https://techcrunch.com/2018/05/24/facebook-political-ad-archive/.

126 *See infra* Section III.D.

India by hosting rumors that led to real-world attacks,[127] Facebook announced a policy in July 2018 under which it would remove false information that is intended to incite violence and other physical harm.[128] Under this new policy, Facebook will remove posts (including text and images) that have been created or shared with the purpose of immediately "contributing to or exacerbating violence or physical harm."[129] To implement the policy, Facebook intends to work with local and international organizations, as well as with its own internal image recognition technologies, to identify such content for removal.[130]

5. Eliminating Fake Accounts

Facebook has also undertaken measures to eliminate fake accounts,[131] i.e., accounts that misrepresent the true identity of the account holder, such as Macedonians pretending to be Americans in order to spread false posts on Facebook.[132] Foreign and domestic operatives have used fake accounts and pages with increasing frequency in past months to make themselves appear more popular.[133] In October 2018, Facebook identified for removal "559 pages and 251 accounts run by Americans, many of which amplified false and misleading content in a coordinated fashion," including Right Wing News, discussed above, and left-wing pages, including the Resistance and Reverb Press, each of which had several hundreds of thousands of followers.[134] These sites targeted for removal amounted to the greatest number of domestic pages and accounts that Facebook has ever removed related to election interference.[135]

127 *See* Max Fisher & Amanda Taub, *In Search of Facebook's Heroes, Finding Only Victims*, N.Y. TIMES (Apr. 22, 2018), https://www.nytimes.com/2018/04/22/insider/facebook-victims-sri-lanka.html?module=inline; Amanda Taub & Max Fisher, *Where Countries Are Tinderboxes and Facebook Is a Match*, N.Y. TIMES (Apr. 21, 2018), https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html?module=inline.

128 *See* Michelle Castillo, *Facebook Will Begin Taking Down Fake News Intended to Encourage Violence*, CNBC (July 18, 2018), https://www.cnbc.com/2018/07/18/facebook-to-take-down-fake-news-intended-to-encourage-violence.html; *The Country Where Facebook Posts Whipped Up Hate*, BBC (Sept. 12, 2018), https://www.bbc.com/news/blogs-trending-45449938 (describing various factors that led to the weaponization of Facebook against Rohingyas in Myanmar). Facebook has already implemented this policy in Sri Lanka to remove posts in Sri Lanka alleging that Muslims were poisoning food that was given or sold to Buddhists. *See* Castillo, *supra*.

129 Castillo, *supra* note 128.

130 *Id.*

131 *See, e.g.*, Tony Romm & Drew Harwell, *Facebook Disabled 583 Million Fake Accounts and Millions of Spam, Sex and Hate Speech Posts*, WASH. POST (May 15, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/05/15/facebook-disabled-583-million-fake-accounts-and-millions-of-spam-sex-and-hate-speech-posts/?utm_term=.e52e7575e7af.

132 *See* Samanth Subramanian, *Inside the Macedonian Fake-News Complex*, WIRED (Feb. 15, 2017), https://www.wired.com/2017/02/veles-macedonia-fake-news.

133 *See* Frenkel, *supra* note 57.

134 *Id.*

135 *Id.*

6. Providing Contextual and Source Information

Facebook is also providing more contextual information about the sources of information made available on its platform by placing an "i" icon on the bottom right of posts that users can click to access information from Wikipedia about the source of the posted information.[136] In addition, Facebook is undertaking a variety of efforts to educate its users on news literacy and to instruct users on how to distinguish between false posts and accurate posts.[137]

7. Allowing Users to Rank Trustworthiness of News Sources

Facebook recently introduced a new scoring system that will allow users to assign news organizations "trust scores," which will be among the factors used to determine how frequently a source appears in a news feed.[138] News organizations that score higher will have their articles more widely distributed.[139] The scoring system's survey will ask a "diverse and representative" sample of users "if they've heard of a news outlet and how much they trust it."[140]

8. Modifying News Feed

Facebook has also modified its news feed functionality in a number of ways and now allows users to customize their news feeds. First, as discussed above, Facebook significantly reduces the news feed rank of each post that fact-checkers have determined to be false, such that the post's future views will be reduced by an average of eighty percent.[141] Second, as of early 2018, Facebook has begun prioritizing interactive content over passive content within its news feed algorithm, such that posts that have received multiple likes, reactions, comments, and shares—especially from a user's closest friends (i.e., those with whom the user interacts the most)—will be priori-

136 *See* Andrew Anker et al., *New Test to Provide Context About Articles*, FACEBOOK NEWSROOM (Oct. 5, 2017), https://newsroom.fb.com/news/2017/10/news-feed-fyi-new-test-to-provide-context-about-articles. *But see* Louise Matsakis, *Don't Ask Wikipedia to Cure the Internet*, WIRED (Mar. 16, 2018), https://www.wired.com/story/youtube-wikipedia-content-moderation-internet.

137 Lyons, *supra* note 14 ("We also continue to invest in news literacy programs to help people better judge the publishers and articles they see on Facebook.").

138 Jacob Kastrenakes, *Facebook Begins Rating Users on How Trustworthy They Are at Flagging Fake News*, VERGE (Aug. 21, 2018), https://www.theverge.com/2018/8/21/17763886/facebook-trust-ratings-fake-news-reporting-score; *see* Adam Mosseri, *Helping Ensure News on Facebook Is from Trusted Sources*, FACEBOOK NEWSROOM (Jan. 19, 2018), https://newsroom.fb.com/news/2018/01/trusted-sources.

139 Mosseri, *supra* note 128.

140 Heather Kelly, *Facebook to Rank News Outlets by Trustworthiness*, CNNMONEY (Jan. 19, 2018), http://money.cnn.com/2018/01/19/technology/facebook-news-trustworthy/index.html?iid=EL.

141 *See* Lyons, *supra* note 14.

tized in the news feed.[142] In another aspect of its news feed modifications, Facebook is now allowing users to customize their news feeds by choosing which posts to see first and/or to hide posts that they do not wish to see.[143]

9. Analysis of Facebook's Efforts

Facebook's efforts to remedy the flaws in the online marketplace of ideas are not subject to First Amendment scrutiny, since it is not a state actor; however, the measures that Facebook has implemented in the United States are largely consistent with the marketplace of ideas theory of the First Amendment and some of these efforts appear to be effective in achieving their goals. First, Facebook's efforts in response to false speech to develop and post "Related Articles"/"Additional Reporting on This" center on providing counterspeech and promoting exposure to diverse and conflicting viewpoints, instead of implementing censorship as a remedy. As discussed above, since the formative years of modern First Amendment jurisprudence, the accepted response to bad speech is not censorship but more speech and

142 *See* Seth Fiegerman & Laurie Segall, *Facebook to Show More Content from Friends, Less from Publishers and Brands*, CNNMONEY (Jan. 11, 2018), http://money.cnn.com/2018/01/11/technology/facebook-news-feed-change/index.html?iid=EL (describing the change as a "'rebalancing' of how Facebook's algorithms rank items in the main feed"); Mark Zuckerberg, FACEBOOK (Jan. 11, 2018), https://www.facebook.com/zuck/posts/10104413015393571?pnref=story ("Facebook has always been about personal connections. By focusing on bringing people closer together—whether it's with family and friends, or around important moments in the world—we can help make sure that Facebook is time well spent."). Facebook based this change in its algorithms to prioritize posts from close friends and family over public content on research showing that using social media to connect with others correlates with greater happiness, health, and well-being, compared to "passively" consuming information from public articles or videos. Fiegerman & Segall, *supra* (explaining why this change is unlikely to impact ads, which operate separately from the news feed system, because businesses can "sidestep this shift" by spending more money on ads to promote their content); Adam Mosseri, *Bringing People Closer Together*, FACEBOOK NEWSROOM (Jan. 11, 2018), https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together ("Because space in News Feed is limited, showing more posts from friends and family and updates that spark conversation means we'll show less public content, including videos and other posts from publishers or businesses. . . . Using 'engagement-bait' to goad people into commenting on posts is not a meaningful interaction, and we will continue to demote these posts in News Feed."). Former Facebook Vice President of Product Adam Mosseri expressed confidence that this change would improve the filter bubble issue because it will showcase conversations with friends over publishers. Fiegerman & Segall, *supra* (quoting Mosseri saying, "You pick a publisher based on your interests, which are more correlated with your beliefs . . . You pick a friend for lots of different reasons. . . . Because this [change] is naturally good for friend content and for conversation, it's actually going to be good for the diversity of opinion in News Feed." (alteration in original)). Other commentators challenge this assumption. For example, CNN's Seth Fiegerman and Laurie Segall claim, "By prioritizing content that sparks conversations, Facebook could risk promoting more polarizing and opinionated posts that generate lots of comments, only adding to the filter bubble." *Id.*

143 *Control What You See in News Feed*, FACEBOOK HELP CTR., https://www.facebook.com/help/964154640320617/?helpref=hc_fnav (last visited Sept. 29, 2018).

more counterspeech. The responses of counterspeech and of facilitating exposure to diverse and conflicting viewpoints were credited by the Supreme Court and implemented by the legislature in the mid- and late-twentieth century in the context of the fairness doctrine for broadcast[144] and the must-carry doctrine for cable.[145] These doctrines were upheld by the Supreme Court as consistent with the First Amendment.[146] The Federal Communications Commission's fairness doctrine imposed on broadcasters the obligation (among others) to afford a reasonable opportunity for discussion of competing points of view and controversial issues of public importance. In upholding the fairness doctrine in *Red Lion Broadcasting Co. v. FCC*, the Supreme Court emphasized "the First Amendment goal of producing an informed public capable of conducting its own affairs" and "the 'public interest' in . . . the presentation of vigorous debate of controversial issues of importance and concern to the public."[147] The must-carry doctrine, which required cable systems operators to carry the signals of local commercial and noncommercial educational public broadcast television stations, was upheld by the Court in *Turner Broadcasting System, Inc. v. FCC*.[148] As the Court explained in upholding the must-carry doctrine, "it has long been a basic tenet of national communications policy that 'the widest possible dissemination of information from diverse and antagonistic sources is essential to the welfare of the public.'"[149] Facebook's efforts to present users with information from diverse and antagonistic sources are consistent with the marketplace of ideas model and the values of counterspeech embodied in this model. These efforts contribute toward "producing an informed public capable of conducting its own affairs" and facilitating the preconditions necessary for citizens to engage in the task of democratic self-government.[150] Facebook's efforts to procure and post "Related Articles"/"Additional Reporting on This" in response to false speech—and to attempt to draw users' attention to such counterspeech and away from false speech—is far more consistent with American free speech values than is the censorial approach mandated by the European Union and by Germany in particular, which requires Facebook to block access to a wide array of speech flagged as illegal by users.

Further, Facebook's efforts to reduce the audience of purveyors of false news and ultimately to remove those purveyors' ability to communicate via its

144 *See* Jerome A. Barron, *The Federal Communications Commission's Fairness Doctrine: An Evaluation*, 30 GEO. WASH. L. REV. 1 (1961).

145 *See* Matthew D. Bunker & Charles N. Davis, *The First Amendment as a Sword: The Positive Liberty Doctrine and Cable Must-Carry Provisions*, 40 J. BROADCASTING & ELECTRONIC MEDIA 77 (1996).

146 Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622 (1994) (cable); Red Lion Broad. Co. v. FCC, 395 U.S. 367 (1969) (broadcast).

147 *Red Lion Broad. Co.*, 395 U.S. at 385, 392.

148 512 U.S. 622.

149 *Id.* at 663–64 (quoting United States v. Midwest Video Corp., 406 U.S. 649, 668 n.27 (1972) (plurality opinion)).

150 *Red Lion Broad. Co.*, 395 U.S. at 392.

platform are generally consistent with the limited protection for false and harmful statements of fact under the Court's defamation jurisprudence.

In addition, Facebook's recently implemented measure of removing false information likely to incite violence is consistent with the emergency exception to the marketplace of ideas model, as originally articulated by Holmes and Brandeis[151] and as recognized by the Court in its incitement jurisprudence in *Brandenburg v. Ohio* and its progeny. The content that Facebook does remove under its new policies—i.e., content that was created or shared with the purpose of immediately contributing to or exacerbating violence or physical harm—is content that would generally be subject to government regulation under the First Amendment's incitement jurisprudence, under which the government is permitted to regulate "advocacy of the use of force or of law violation . . . where such advocacy is directed to inciting or producing imminent lawless action and is likely to incite or produce such action."[152]

Other efforts undertaken by Facebook to remedy the flaws in the online marketplace of ideas appear less likely to comport with First Amendment values and less likely to be effective in fixing the flaws in the online marketplace of ideas. Measures like allowing people to customize their news feeds to prioritize posts from friends and family over public content—as well as allowing users to rank the trustworthiness of news sources—seem likely to entrench information silos and filter bubbles. Such measures, which are likely to limit users exposure to content from a diverse array of sources, run counter to the important First Amendment goals of promoting "the widest possible dissemination of information from diverse and antagonistic sources" to advance the welfare of the public.[153]

As one of the most important forums for expression in the United States, Facebook should continue to focus on implementing remedies to fix the flaws in the online marketplace of ideas that are consistent with First Amendment values, including its recent measures focusing on counterspeech instead of censorship as a response to false speech.

Recent empirical studies suggest that Facebook's efforts to combat false news have been moderately successful. As Hunt Allcott, Matthew Gentzkow, and Chuan Yu report in the October 2018 article *Trends in the Diffusion of Misinformation on Social Media*, based on their study of "trends in the diffusion of content from 570 fake news websites and 10,240 fake news stories on Facebook and Twitter between January 2015 and July 2018," while "[u]ser interactions with false content rose steadily on . . . Facebook . . . through the end of 2016," since then, "interactions with false content have fallen sharply."[154] The authors of the study find that "user interaction with known

151 *See supra* notes 30–31 and accompanying text.

152 Brandenburg v. Ohio, 395 U.S. 444, 447 (1969) (per curiam).

153 *Turner Broad. Sys., Inc.*, 512 U.S. at 663–64 (quoting *Midwest Video Corp.*, 406 U.S. at 668 n.27).

154 Allcott et al., *supra* note 101, at 1.

false news sites has declined by 50 percent since the 2016 election."[155] Based on these findings, the authors conclude that "efforts by Facebook following the 2016 election to limit the diffusion of misinformation [namely, the "suite of policy and algorithmic changes made by Facebook following the election"] may have had a meaningful impact."[156] In light of these findings, Facebook should work to identify which of its efforts have been most successful in limiting user interaction with known false news sites and should redouble such efforts.

C. What Twitter Is Doing

In addition to complying with the mandates from the European Union—under the EU Code of Conduct and the new EU Code of Practice on Disinformation—and from Germany under its NetzDG legislation, and in addition to removing content that violates its own terms of service,[157] Twitter has implemented a number of measures to attempt to remedy the flaws in the online marketplace of ideas, and is contemplating the implementation of additional measures. Twitter's response to the flaws in the online marketplace of ideas—and especially to the issues caused by Russian operatives' use of fake Twitter accounts on a massive scale to interfere with U.S. elections—has been largely focused on suspending fake and suspicious accounts,[158] and has recently embraced regulation of political advertising as well. Twitter has also modified its algorithms to prioritize the highest quality and most relevant content.

1. Suspending Fake and Suspicious Accounts

In the wake of congressional pressure, Twitter has sharply escalated its efforts to remove fake and suspicious accounts by suspending over one million accounts per day in recent months in an attempt to limit the spread of false news and misinformation via its platform.[159] Indeed, in May and June 2018 alone, Twitter suspended over seventy million accounts.[160] This aggressive campaign against bots is a reaction to Russia's social media disinforma-

155 Fowler, *supra* note 114; *see* Allcott et al., *supra* note 101, at 5.

156 Allcott et al., *supra* note 101, at 3, 6.

157 The "Twitter Rules," which are included in the "Twitter User Agreement," prevent users from posting the following types of content, as set forth more fully in Appendix B: content that violates others' intellectual property rights, graphic violence and adult content, distribution of hacked materials, misuse of usernames, abusive behavior, violence and physical harm, suicide or self-harm, child sexual exploitation, unwanted sexual advances, abuse and hateful conduct, hateful imagery and display names, private information, intimate media, threats to expose/hack, and impersonation. *The Twitter Rules*, Twitter Help Ctr., https://help.twitter.com/en/rules-and-policies/twitter-rules (last visited Dec. 3, 2018).

158 *See* Timberg & Dwoskin, *supra* note 54; *see also* Allcott et al., *supra* note 101, at 5 (detailing a smaller set of initiatives to combat false news compared to Facebook's efforts).

159 *See* Timberg & Dwoskin, *supra* note 54.

160 *See id.*

tion campaign during the 2016 elections, particularly with regard to the St. Petersburg–based troll factory known as the Internet Research Agency.[161]

Recently, Twitter has also been responding to efforts by domestic operatives to mimic the "success" of foreign operatives by flooding the Twitter platform with misinformation. In October 2018, Twitter took down a network of fifty accounts that were being run by Americans posing as Republican state lawmakers and that were targeting voters in all fifty states.[162] In addition, in the two-month period leading up to the 2018 midterm elections, Twitter deleted more than ten thousand automated accounts posting messages that discouraged people from voting in the election and that falsely appeared to come from Democrats.[163] Twitter has also recently modified its policies governing the removal of fake accounts. As announced in an October 1, 2018, blog post entitled *An Update on Our Elections Integrity Work*, Twitter has expanded its ability to remove fake accounts.[164] Twitter explained:

> [W]e are updating and expanding our rules to better reflect how we identify fake accounts, and what types of inauthentic activity violate our guidelines. We now may remove fake accounts engaged in a variety of emergent, malicious behaviors. Some of the factors that we will take into account when determining whether an account is fake include:
>
> –Use of stock or stolen avatar photos
>
> –Use of stolen or copied profile bios
>
> –Use of intentionally misleading profile information, including profile location.[165]

In this post, Twitter further explained:

> [I]f we are able to reliably attribute an account on Twitter to an entity known to violate the Twitter Rules, we will take action on additional accounts associated with that entity. We are expanding our enforcement approach to include accounts that deliberately mimic or are intended to replace accounts we have previously suspended for violating our rules.[166]

---

161 *See id.*

162 Frenkel, *supra* note 57.

163 Christopher Bing, *Exclusive: Twitter Deletes over 10,000 Accounts That Sought to Discourage U.S. Voting*, REUTERS (Nov. 2, 2018), https://www.reuters.com/article/us-usa-election-twitter-exclusive/exclusive-twitter-deletes-over-10000-accounts-seeking-to-discourage-voting-idUSKCN1N72FA; Ben Collins, *In Secret Chats, Trolls Struggle to Get Twitter Disinformation Campaigns off the Ground*, NBC NEWS (Nov. 5, 2018), https://www.nbcnews.com/tech/tech-news/secret-chats-trolls-struggle-get-twitter-disinformation-campaigns-ground-n931756.

164 Del Harvey & Yoel Roth, *An Update on Our Elections Integrity Work*, TWITTER BLOG (Oct. 1, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/an-update-on-our-elections-integrity-work.html.

165 *Id.*

166 *Id.*

2. Mechanisms for User Reporting of Content and Accounts

In contrast to Facebook, Twitter does not provide a mechanism for users to trigger a fact-checking inquiry in the veracity of posts.[167] Twitter does provide a tool for viewers to report content[168] and/or to report entire accounts. In connection with each tweet, Twitter allows users to click "Report Tweet," which provides the following options: (1) "I'm not interested in this Tweet," (2) "It's suspicious or spam," (3) "It's abusive or harmful."[169] Twitter also allows users to report entire accounts as being "abusive or harmful."[170] In addition, Twitter allows users to block and/or mute accounts that they do not wish to view tweets from.[171]

3. Demoting Tweets from Bad Faith Actors

Twitter is also employing the measure of limiting the reach of tweets from "bad-faith actors who intend to manipulate or divide . . . healthy public conversation" by placing such tweets lower down in the stream of messages in a user's Twitter feed.[172] It has modified its algorithms in an attempt to prioritize the "highest quality and most relevant content,"[173] and to deprioritize tweets from "bad-faith actors."[174]

4. Transparency and Disclosure Requirements re Political/Electioneering Advertisements

Twitter has also implemented new policies regarding political advertisements to prohibit foreign operatives from purchasing political ads and to enhance transparency for such ads.[175] In October 2017, Twitter announced that it intended to implement a political campaigning policy to increase transparency regarding all advertisements on its platform, including politi-

167 *See* Rob Price, *Twitter Says It Doesn't Plan to Launch a Button to Report Fake News*, BUSINESS INSIDER (June 30, 2017), https://www.businessinsider.com/twitter-not-planning-launch-button-report-fake-news-washington-post-report-2017-6.

168 *See* Alex Murray, *How to Report Fake News to Social Media*, BBC (Nov. 22, 2016), https://www.bbc.com/news/38053324.

169 *See* TWITTER, http://twitter.com (last visited Mar. 14, 2019).

170 *See Report Abusive Behavior,* TWITTER HELP CTR., https://help.twitter.com/en/safety-and-security/report-abusive-behavior (last visited Dec. 14, 2018).

171 *See Using Twitter*, TWITTER HELP CTR., https://help.twitter.com/en/using-twitter#blocking-and-muting (last visited Dec. 14, 2018).

172 Vijaya Gadde & Kayvon Beykpour, *Setting the Record Straight on Shadow Banning*, TWITTER BLOG (July 26, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html.

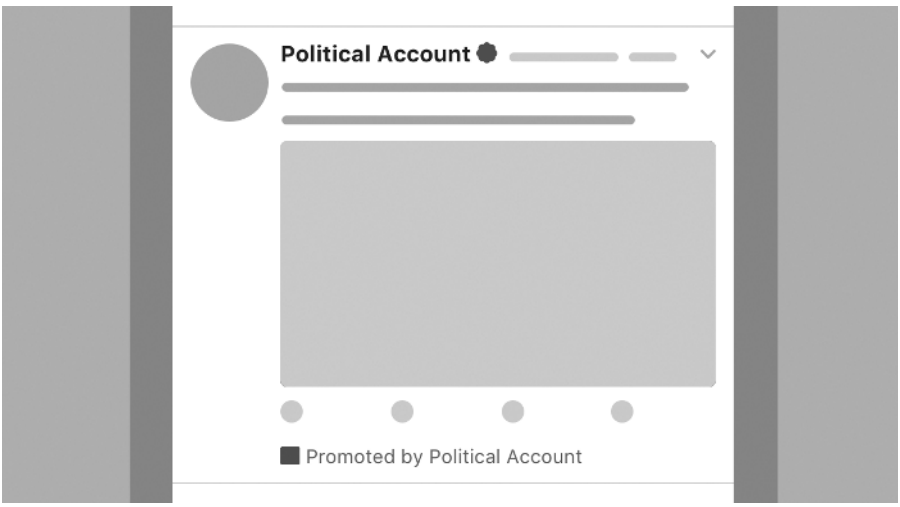
173 *See* Colin Crowell, *Our Approach to Bots and Misinformation*, TWITTER BLOG (June 14, 2017), https://blog.twitter.com/official/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html.

174 Gadde & Beykpour, *supra* note 172.

175 *See* Vijaya Gadde & Bruce Falck, *Increasing Transparency for Political Campaigning Ads on Twitter*, TWITTER BLOG (May 24, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/Increasing-Transparency-for-Political-Campaigning-Ads-on-Twitter.html.

cal/electioneering ads and issue-based ads.[176] Twitter implemented these policies in June 2018.[177] As part of its new political campaigning policy,[178] Twitter requires all "advertisers who want to run political campaigning ads for [U.S.] Federal elections to self-identify and certify that they are located in the U.S."[179] Under this policy, Twitter prohibits foreign nationals from targeting political ads "to people who are identified as being in the U.S."[180] As part of these recently implemented measures, Twitter also now includes a "visual political ad indicator" (see Figure 3) for all "electioneering" ads—those that refer to a clearly identified candidate (or party associated with that candidate) for any elected office—and requires electioneering advertisers to identify their campaigns.[181]

FIGURE 3



Twitter has also set up an "Ads Transparency Center" that includes, for each electioneering ad, (1) "[d]isclosure on total campaign ad spend by advertiser"; (2) "[t]ransparency about the identity of the organization funding the campaign"; (3) "[t]argeting demographics, such as age, gender and geography"; and (4) "[h]istorical data about all electioneering ad spending

176 *See id.*

177 *See* Bruce Falck, *Providing More Transparency Around Advertising on Twitter*, TWITTER BLOG (June 28, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/Providing-More-Transparency-Around-Advertising-on-Twitter.html.

178 *See Political Content*, TWITTER BUS., https://business.twitter.com/en/help/ads-policies/restricted-content-policies/political-campaigning.html (last visited Dec. 14, 2018).

179 *See* Gadde & Falck, *supra* note 175.

180 *Id.*

181 *See* Bruce Falck, *New Transparency for Ads on Twitter*, TWITTER BLOG (Oct. 24, 2017) https://blog.twitter.com/official/en_us/topics/product/2017/New-Transparency-For-Ads-on-Twitter.html.

by advertiser."[182] In addition, Twitter revised its policies for electioneering advertisers to (1) "[i]nclude stricter requirements on who can serve these ads and to limit targeting options," (2) "[r]equire electioneering advertisers to self-identify as such," and (3) provide "stronger penalties for advertisers who violate these policies."[183] Indeed, Twitter's recently implemented measures limiting foreign entities from purchasing political ads and imposing disclosure requirements on political ads go beyond those that are encompassed in the proposed Honest Ads Act, discussed below, and manifest a commitment from Twitter to meaningfully address the problems of foreign interference in the U.S. political process.

5. Future Plans

Twitter's CEO, Jack Dorsey, has also promised to develop and implement future plans to combat false news and create a healthier discourse, including "rethinking core parts of Twitter to curb the spread of hate speech, harassment and false news."[184] Dorsey has also indicated that Twitter intends to experiment with the implementation of "features that would allow people to see alternative viewpoints and reduce 'echo chambers.'"[185]

6. Analysis of Twitter's Efforts

Twitter's efforts to combat false news on its platform have been less extensive than Facebook's comparable efforts, and have largely focused on removing fake and suspicious accounts. Recent empirical studies indicate that Twitter's efforts to combat false news have been less successful than the efforts undertaken by Facebook.[186] According to a recent analysis by Matthew Hindman and Vlad Barash, as of October 2018, Twitter was still awash with fake news, with more than eighty percent of the accounts that regularly spread misinformation in 2016 still active.[187] The report examined more than seven hundred thousand Twitter accounts, which linked to more than six hundred sites spreading misinformation or conspiracy stories, and found that those accounts still publish over one million tweets per day.[188] As dis-

182 *Ads Transparency Center*, TWITTER, https://ads.twitter.com/transparency (last visited Mar. 15, 2019).

183 Flack, *supra* note 181.

184 Kristine Phillips, *Twitter CEO Jack Dorsey Admits 'Left-Leaning' Bias but Says It Doesn't Influence Company Policy*, WASH. POST (Aug. 19, 2018), https://www.washingtonpost.com/technology/2018/08/19/twitter-ceo-jack-dorsey-admits-left-leaning-bias-says-it-doesnt-influence-company-policy/?utm_term=.282e9bf8732b.

185 *See id.*; CNN, *Twitter CEO: 'We Are Not' Discriminating Against Any Political Viewpoint*, YOUTUBE (Aug.19, 2018), https://www.youtube.com/watch?v=Cm_lmWWKDug.

186 *See* Allcott et al., *supra* note 101, at 5 (observing that fake news interactions on Facebook fell sharply following the 2016 U.S. presidential election—declining by more than fifty percent—while Twitter shares continued to increase).

187 Jason Schwartz, *Twitter Still Awash in Fake News, Study Finds*, POLITICO (Oct. 4, 2018), https://www.politico.com/story/2018/10/04/twitter-fake-news-866676.

188 *Id.*

cussed above, unlike Facebook, Twitter has not implemented a mechanism for users to flag false news or false accounts, nor has it worked with third-party fact-checkers to evaluate the veracity of posts or to develop counter-speech in response to posts determined to be false. Twitter should adopt and implement measures similar to those that Facebook has implemented that have been shown to be most effective at combatting false news and that are consistent with First Amendment values.

D. What the U.S. Legislature Seeks to Do: The Honest Ads Act

The proliferation of false statements, misleading information, and fake accounts in the online marketplace of ideas—predominantly originated by foreign sources—has had a profoundly harmful effect on our democratic process, as discussed above. Yet, social media platforms have been largely immune from federal election campaign regulations and related regulations that have long been applicable to other sources of news and information in our political information ecosystem that mandate transparency and accountability requirements.[189] Various federal statutes, Federal Election Commission rules,[190] and Federal Communications Commission rules[191] currently impose transparency requirements on political advertisements disseminated by broadcast, cable, and satellite providers, and also impose requirements on these providers prohibiting foreign participation in U.S. elections; yet, online platforms like Facebook and Twitter are currently not subject to analogous regulations (although these platforms have recently committed to self-regulation in this arena, as discussed above[192]). This is despite the fact that Facebook's user base of 204 million American users is ten times larger than the subscriber base of the largest cable and satellite providers[193] and despite the fact that over one billion dollars was spent on online advertising in 2016.[194] This lack of regulation has allowed foreign actors to influence the electorate, including by allowing Russian entities to purchase approximately three thousand ads between June 2015 and May 2017 linked to fake accounts

189 *See infra* notes 190–92 and accompanying text.

190 *See* Fredreka Schouten, *Federal Regulators Approve Narrow Facebook Ad Disclosure*, USA TODAY, (Dec. 14, 2017), https://www.usatoday.com/story/news/politics/2017/12/14/federal-regulators-weigh-whether-unmask-online-political-ad-buyers/951425001.

191 *E.g.*, 47 U.S.C. § 317 (2012).

192 *See supra* Section III.B, III.C.

193 *Compare Leading Countries Based on Number of Facebook Users as of July 2018 (in Millions)*, STATISTA, https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users (last visited Sept. 29, 2018) (showing the amount of U.S. Facebook users), *with Pay TV Providers Ranked by the Number of Subscribers in the United States as of September 2018 (in Millions)*, STATISTA, https://www.statista.com/statistics/251793/pay-tv-providers-with-the-largest-number-of-subscribers-in-the-us/ (last visited Jan. 23, 2019) (indicating that the top television provider in the United States has only twenty million subscribers).

194 *Digital Political Advertising Spending in the United States from 2008 to 2020 (in Million U.S. Dollars)*, STATISTA, https://www.statista.com/statistics/309592/online-political-ad-spend-usa/ (last visited Jan. 31, 2019).

associated with the pro-Kremlin Internet Research Agency.[195] The Honest Ads Act, introduced in October 2017 by Senators Mark Warner (D-VA), Amy Klobuchar (D-MN) and the late John McCain (R-AZ), seeks to remedy this regulatory disparity. The Act attempts to address some of the problems created by foreign interference in U.S. elections in the online arena by imposing transparency regulations on online political advertisements and by requiring that online platforms enforce the longstanding ban on foreign participation in U.S. elections.[196] Although, as discussed above, social media platforms like Facebook and Twitter are undertaking substantial measures themselves to address such problems, government regulation in the form of the Honest Ads Act is also an important tool to address these problems, and one welcomed by the platforms.[197]

The Honest Ads Act seeks to address problems in the online marketplace of ideas by extending three sets of requirements that have long been imposed on communications platforms to online platforms: (1) the expansion of disclosure requirements applicable to political advertisements; (2) the expansion of public file requirements; and (3) the expansion of the obligation to undertake reasonable efforts to limit foreign interference in U.S. elections.

First, the Honest Ads Act extends the disclosure obligations governing political advertisements that print, broadcast, and cable advertisements must meet to online platforms.[198] The Federal Election Campaign Act of 1971 requires that certain political ads in print, broadcast, and cable disclose who has paid for the advertisement.[199] This requirement currently does not extend to paid internet or digital advertisements. Under the Honest Ads Act, the Federal Election Campaign Act's definition of "electioneering communication" would be expanded to include online paid political advertisements.[200] Existing federal law also imposes disclosure requirements on "public communications" that expressly advocate for a candidate's election or defeat, are paid for or authorized by a candidate, solicit a political contribution, or are made by a political committee.[201] The Honest Ads Act

195 *See* Schouten, *supra* note 190.

196 *See* Honest Ads Act, S. 1989, 115th Cong. §§ 3–4 (2017).

197 Both Facebook and Twitter have come out in support of the Honest Ads Act. *See* Aimee Picchi, *Facebook: What Is the Honest Ads Act?*, CBS NEWS, (Apr. 11, 2018), https://www.cbsnews.com/news/facebook-hearings-what-is-the-honest-ads-act; Twitter Public Policy (@Policy), TWITTER (Apr. 10, 2018, 11:54 AM), https://twitter.com/Policy/status/983734917015199744.

198 S. 1989, § 2 ("The purpose of this Act is to enhance the integrity of American democracy and national security by improving disclosure requirements for online political advertisements . . . .").

199 52 U.S.C. § 30120 (2012).

200 S. 1989, § 6.

201 *Advertising and Disclaimers*, FED. ELECTION COMMISSION, https://www.fec.gov/help-candidates-and-committees/making-disbursements/advertising/ (last visited Mar. 15, 2019); *Public Communications*, FED. ELECTION COMMISSION, https://www.fec.gov/press/resources-journalists/public-communications/ (last visited Mar. 15, 2019).

updates the definition of "public communication" as well, to ensure that disclosure obligations applicable to these types of advertisements extend to the online environment as well.[202]

Second, under the Honest Ads Act, large digital platforms (those with more than fifty million unique monthly visitors) would be required to maintain publicly available records of political advertisements by a purchaser whose aggregate requests to purchase political advertisements on that platform exceed $500 within the past year.[203] Such records must include a digital copy of the political advertisement, as well as a description of the target audience, the ad rate, the name of the candidate or office that the ad was supporting, and the contact information of the purchaser of the ad.[204] Like the Federal Communication Commission's broadcast file rules, the Act would apply to ads made by, for, or about political candidates, about elections, and about "national legislative issue[s] of public importance."[205]

Third, the Honest Ads Act would mandate that all advertising platforms—including online platforms—make reasonable efforts to comply with the foreign participation ban.[206] This longstanding ban prohibits foreign nationals from attempting to influence elections through donations, expenditures, or other things of value.[207] Existing regulations applicable to broadcast, cable, and satellite platforms include a broad prohibition on the involvement of foreign nationals with elections in the United States, under which foreign nationals are prohibited from making any contribution, donation, or expenditure in connection with any federal, state, or local election; making any contribution or donation to any committee or organization of any national, state, or local political party; or making any disbursement for an electioneering communication. The Act would extend these prohibitions to the online environment as well.

The Honest Ads Act's extension of regulations regarding elections and election-related advertisements to online platforms is an important step toward fixing the flaws in the online marketplace of ideas and is consistent with First Amendment law. Since the early days of campaign finance regulation the Supreme Court has upheld legislative efforts aimed at advancing transparency and facilitating other conditions necessary for an informed electorate, and the Honest Ads Act's extension of requirements and obligations to online platforms to advance these goals is likely to be upheld against First Amendment challenge. Even the Supreme Court's much maligned *Citizens United v. FEC* decision—which struck down most of the statutory provisions of the Bipartisan Campaign Reform Act (BCRA)—upheld the Act's disclosure and public file requirements, which were aimed at informing the

202 S. 1989, § 5.

203 *Id.* § 8(a).

204 *Id.*

205 *Id.*

206 *See id.* § 9.

207 52 U.S.C. § 30121(a)(1)(A) (2012).

electorate about the source of election-related advertisements.[208] Among its other challenges to BCRA, Citizens United brought a challenge to the BCRA's disclosure provisions as applied to its movie *Hillary* and to three television advertisements for the movie.[209] These ads fell within BCRA's definition of "electioneering communication" because they referred to candidate Hillary Clinton by name shortly before a primary election and contained pejorative references to her candidacy.[210] BCRA section 311 requires that televised "electioneering communications" funded by anyone other than a candidate must include a statement clearly indicating who is responsible for the ad (in the form of "is responsible for the content of this advertising"),[211] along with the name and address (or web address) of the person who funded the ad.[212] In addition, BCRA requires that anyone who spends more than $10,000 on electioneering communications within a calendar year file a detailed statement with the FEC, providing his or her name, amount of expenditure, and the name of the election to which the communication is directed, among other details.[213] Citizens United challenged the disclosure and public file provisions of BCRA, arguing that they unconstitutionally burdened its First Amendment rights.[214] The Supreme Court rejected these challenges and upheld these requirements. The Court explained that these provisions "'provid[e] the electorate with information' . . . and 'insure that the voters are fully informed' about the person or group who is speaking' . . . 'so that people will be able to evaluate the arguments to which they are being subjected.'"[215] The Court concluded that these requirements were a less restrictive alternative compared to other, more extensive regulations of campaign speech, that "the public has an interest in knowing who is speaking about a candidate shortly before an election," and that this "informational interest alone is sufficient to justify application of [the Act] to these ads."[216]

In summary, the Honest Ads Act's provisions—which extend existing disclosure and public file requirements to online platforms and require that such platforms comply with the foreign participation ban—would substantially address flaws in the online marketplace of ideas that were prevalent in connection with the 2016 U.S. presidential election and thereafter, and should be adopted by the legislature and upheld by the courts as consistent with the First Amendment.

---

208 Citizens United v. FEC, 558 U.S. 310, 368–71 (2010).

209 *Id.* at 319–21.

210 *Id.* at 322–23.

211 52 U.S.C. § 30120(d)(2).

212 *Id.* § 30120(a)(3).

213 *Id.* §§ 30104(f)(1)–(2).

214 Citizens United v. FEC, 558 U.S. 310, 321 (2010).

215 *Id.* at 368 (citations omitted) (first quoting McConnell v. FEC, 540 U.S. 93, 196 (2003); then quoting Buckley v. Valeo, 424 U.S. 1, 76 (1976) (per curiam); and then quoting First Nat'l Bank of Bos. v. Bellotti, 435 U.S. 765, 792 n.32 (1978)).

216 *Id.* at 369.

E. What U.S. Litigants Are Doing: Defamation Actions Against Purveyors of False News

Another significant way in which flaws in the online marketplace of ideas are being addressed is through the vehicle of defamation lawsuits brought by targets of false news against those who traffic in and profit from such content. Prominent recent examples of defamation actions against false news purveyors include those recently brought by the families of the Sandy Hook victims against notorious InfoWars publisher Alex Jones.[217] Jones's InfoWars website has published several articles and videos claiming that the December 2012 Sandy Hook massacre never actually happened and was instead an elaborate hoax invented by government-backed "gun grabbers" to limit Americans' Second Amendment rights.[218] Jones maintained that the Sandy Hook massacre was staged by the government using actors, and that the family members of the Sandy Hook victims are actually "crisis actors" who are lying about their children being killed.[219] In reliance on the false claims, followers of Jones have stalked and harassed the families of Sandy Hook victims to demand "proof" that their loved ones were actually killed, and have communicated death threats to them, causing the victims to move to protect their families.[220] In three separate lawsuits, the families of Sandy Hook victims, along with an FBI agent who was involved in responding to the massacre, claim that Jones defamed them by accusing them of participating in such fraudulent or illegal activities. The plaintiffs in the lawsuits assert that such claims have harmed their reputations and subjected them to public con-

217 *See, e.g.*, Plaintiff's Original Petition & Request for Disclosure at 14–16, Heslin v. Jones, No. D-1-GN-18-001835 (Tex. Dist. Ct. Apr. 16, 2018); Sebastian Murdock, *Sandy Hook Parents Hit Alex Jones with Defamation Lawsuits*, HUFFINGTON POST, (Apr. 17, 2018, 5:46 AM), https://www.huffingtonpost.com/entry/sandy-hook-parents-hit-alex-jones-with-defamation-lawsuits_us_5acf6a6de4b0ac383d74bfe1. Jones is also facing a defamation lawsuit brought by Brennan Gilmore, the man who recorded the deadly car attack at Charlottesville's white supremacy rally in 2017 that killed Heather Heyer and injured nineteen other people. Rachel Weiner & Abby Ohlheiser, *The Conspiracy Theorists Attacked Him. He's Fighting Back in Court.*, WASH. POST (Mar. 13, 2018), https://www.washingtonpost.com/local/public-safety/witness-sues-infowars-for-claiming-he-caused-charlottesville-protesters-death/2018/03/13/4af0b4ee-26ca-11e8-b79d-f3d931db7f68_story.html?utm_term=.d0e4a2f561f6. Jones falsely called Gilmore a "deep state shill" and a "CIA asset" and falsely accused him of helping to organize the attack as a way of discrediting President Trump and his supporters. Murdock, *supra*. As a result of these false claims, Gilmore has faced death threats, doxxing, and harassment. *See* Brennan Gilmore, *How I Became Fake News: I Witnessed a Terrorist Attack in Charlottesville. Then the Conspiracy Theories Began.*, POLITICO MAG. (Aug. 21, 2017), https://www.politico.com/magazine/story/2017/08/21/fake-news-charlottesville-215514.

218 Elizabeth Williamson, *Truth in a Post-Truth Era: Sandy Hook Families Sue Alex Jones, Conspiracy Theorist*, N.Y. TIMES (May 23, 2018), https://www.nytimes.com/2018/05/23/us/politics/alex-jones-trump-sandy-hook.html.

219 *See id.*

220 *Id.*

tempt, disgrace, ridicule, and attack, and are seeking damages, including punitive damages, upward of one million dollars.[221]

In ruling on such claims, courts are called upon to balance plaintiffs' right to meaningful redress for damage to their reputation and dignitary interests against the First Amendment mandate that debate on public issues be "uninhibited, robust, and wide-open."[222] One of the defamation lawsuits, *Pozner v. Jones*, was brought by Leonard Pozner and Veronique De La Rosa, whose six-year-old son, Noah, was among twenty students and six adults killed at Sandy Hook Elementary School in Newtown, Connecticut, on December 14, 2012.[223] In his motion to dismiss, Jones interposed the First Amendment as well as the Texas Citizens Participation Act, which protects citizens' free speech rights against frivolous lawsuits.[224] Jones also claimed that plaintiffs are public figures (or at least limited-purpose public figures) because Pozner has started a nonprofit to fight against "cruelty and criminality of abusive activity" suffered by victims of tragedies, and De La Rosa has spoken publicly in favor of an assault weapon ban.[225] Because of these activities, Jones claims, these plaintiffs should be subject to a heightened First Amendment burden under the Supreme Court's defamation jurisprudence and should be required to prove that the defendant acted with knowledge that his statements were false or with reckless disregard as to the truth or falsity of his statements.[226] On August 30, 2018, the court denied Jones's motion to dismiss, rejecting Jones's argument that he was entitled to dismissal under the Texas Citizens Participation Act and under the First Amendment.[227]

The defamation action against Jones will now proceed, requiring the court (and jury) to consider whether Jones's false characterizations of the Sandy Hook massacre and the victims' role in the aftermath constitute

221 *Id.*; *see, e.g.*, Plaintiff's Original Petition & Request for Disclosure at 14–17, Pozner v. Jones, No. D-1-GN-18-001842 (Tex. Dist. Ct. Apr. 16, 2018).

222 New York Times Co. v. Sullivan, 376 U.S. 254, 270 (1964).

223 Plaintiff's Original Petition & Request for Disclosure, *supra* note 221, at 2; Williamson, *supra* note 218.

224 Defendant's Motion to Dismiss at 19–34, *Pozner*, No. D-1-GN-18-001842. Under the Texas Citizens Participation Act—Texas's anti-SLAPP legislation—a defendant may move to dismiss a lawsuit by establishing by a preponderance of evidence that the suit is "based on, relates to, or is in response to [the defendant's] exercise of the right of free speech, right to petition, or right of association." TEX. CIV. PRAC. & REM. CODE ANN. § 27.003(a) (West 2017). If the defendant meets his burden, the case must be dismissed unless the plaintiff presents clear and specific evidence of each element of his or her claims, in which case the burden shifts back to the defendant by proving each element of a valid defense.

225 Defendant's Motion to Dismiss, *supra* note 224, at 6–14 (quoting HONR NETWORK, https://www.honr.com/ (last visited Jan. 23, 2019)).

226 *Id.* at 45.

227 *See* Tom Kludt, *Alex Jones' Bid to Throw Out Sandy Hook Defamation Lawsuit Denied*, CNNMONEY (Aug. 31, 2018), https://money.cnn.com/2018/08/30/media/alex-jones-pozner-defamation-suit/index.html (reporting that the court denied Jones's motion to dismiss in all respects); Jorge L. Ortiz, *Alex Jones Denied in Request for Dismissal of Defamation Lawsuit Against Him*, USA TODAY, (Aug. 30, 2018), https://www.usatoday.com/story/news/2018/08/30/alex-jones-request-dismissal-defamation-lawsuit-denied/1150498002.

actionable defamation or protected speech. Although all opinions and ideas are protected by the First Amendment, false statements of fact that harm an individual's reputation are not.[228] While the Supreme Court has—consistent with the marketplace of ideas model—emphasized that there is no such thing as a false *idea*, it has not extended similarly broad immunity to statements that are capable of verification or falsification.[229] Even if the court determines in *Pozner v. Jones* and similar defamation suits that a heightened burden of proof is applicable and requires defamation plaintiffs to establish that Jones made such statements with knowledge of the statements' falsity or reckless disregard of whether the statements were false, plaintiffs should nonetheless prevail on their defamation claims under this standard. Jones and other purveyors of harmful false news should not be able to wield the First Amendment as a defense in cases like this one where he repeatedly profits from false and damaging statements about plaintiffs. In short, courts should not prevent plaintiffs from wielding defamation law as a remedy to help address the real harms and problems caused by false news in the online marketplace of ideas.[230]

CONCLUSION

Today's marketplace of ideas suffers from a host of serious problems that Holmes could never have anticipated when he championed this model one hundred years ago. Fortunately, in adopting Holmes's marketplace model, the Supreme Court has done so in a manner that affords the government sufficient—albeit limited—powers to intervene to remedy flaws in the marketplace of ideas online. Such government intervention is now necessary

228 *E.g.*, New York Times Co. v. Sullivan, 376 U.S. 254 (1964).

229 *See, e.g.*, Gertz v. Robert Welch, Inc., 418 U.S. 323, 340 (1974) ("[T]here is no constitutional value in false statements of fact.").

230 In addition, in order to wield defamation law to meaningfully fix the flaws in the online marketplace of ideas, plaintiffs need to be able to know who is responsible for defamatory content, and courts should not impose unduly burdensome obstacles on plaintiffs in their attempts to uncover the identity of those who anonymously defame them online. Since section 509 of the Communications Decency Act of 1996 insulates web forums from liability for defamatory content of their users, 47 U.S.C. § 230(c) (2012), and since many defamatory statements are posted anonymously, in many cases it is difficult for plaintiffs to identify the individual or individuals who are responsible for statements that allegedly defame them. Courts should not impose overly burdensome requirements on defamation plaintiffs who are seeking to require web forums to disclose the identity of those responsible for allegedly defamatory statements. As one court explained, "[t]hose who suffer damages as a result of tortious or other actionable communications on the Internet should be able to seek appropriate redress by preventing the wrongdoers from hiding behind an illusory shield of purported First Amendment rights." Cohen v. Google, Inc., 887 N.Y.S.2d 424, 425, 429–30 (Sup. Ct. 2009) (quoting *In re* Subpoena Duces Tecum to Am. Online, Inc., 52 Va. Cir. 26, 35 (Cir. Ct. 2000), *rev'd sub nom on other grounds*, Am. Online, Inc. v. Anonymous Publicly Traded Co., 542 S.E.2d 377 (Va. 2001)) (granting defamation plaintiff's petition to compel preaction disclosure requiring Google to turn over information on the identity of an anonymous blogger who allegedly defamed plaintiff by calling her a "skank," "ho," and accused her of "whoring").

to supplement the measures that leading online platforms like Facebook and Twitter are undertaking in an attempt to address such problems. The proposed Honest Ads Act, which extends the obligation to limit foreign intervention in U.S. elections and transparency and accountability obligations regarding political advertisements to online platforms is one such effort that is necessary to fix such flaws in a manner that is consistent with First Amendment law. In addition, online platforms should continue to implement measures to address flaws in the online marketplace of ideas in a manner that advances First Amendment values—including by focusing on counterspeech remedies instead of censorship remedies in response to harmful speech. Finally, the common law of defamation should continue to be invoked by injured parties to address flaws in the online marketplace of ideas, and courts should not invoke the First Amendment to bar such efforts. A combination of new government regulation, use of existing common-law remedies, and self-regulation by the online platforms themselves will enable us to make meaningful progress toward fixing the flaws in the marketplace of ideas online.

APPENDIX A

Facebook Community Standards (Synopsis)

I. Violence and Criminal Behavior
   1. Credible Violence

. . . .

Do not post:

The following threats:

- Credible statements of intent to commit violence against any person, groups of people, or place (city or smaller). We assess credibility based upon the information available to us and generally consider statements credible if the following are present:
  - A target (person, group of people, or place) and
    - Bounty/demand for payment, or
    - Mention or image of specific weapon, or
    - Sales offer or ask to purchase weapon, or
    - Spelled-out address or named building, or
  - A target and two or more of the following details (can be two of the same detail):
    - Location
    - Timing
    - Method
- Any statement of intent to commit violence against a vulnerable person (identified by name, title, image, or other reference) or vulnerable group, including (but not limited to) heads-of-state, witnesses and confidential informants, activists, and journalists

Calls for violence or statements advocating violence against the following targets (identified by name, title, image, or other reference)

- Any vulnerable person or group including (but not limited to) heads of state, national elected officials, witnesses and confidential informants, activists, and journalists
- Public individuals, if credible as defined above
- Groups of people or unnamed specific person(s), if credible
- Places, if credible
- Where no target is specified but a symbol representing the target or a visual of weapons is included

Aspirational and conditional statements of violence against

- Any vulnerable groups
- Public individuals, if credible (unless the individual is convicted of certain crimes or is a member of a dangerous organization)
- Vulnerable person(s), if credible
- Groups of people or unnamed specific person(s), if credible
- Places, if credible

Any content created for the express purpose of outing an individual as a member of a designated and recognizable at-risk group

Instructions on how to make or use weapons if the goal is to injure or kill people as may be evident from:

- As evident from language explicitly stating that goal, or
- As evident from imagery that shows or simulates the end result (serious injury or death) as part of the instruction
- Unless there is clear context that the content is for an alternative purpose (for example, shared as part of recreational self-defense activities, training by a country's military, commercial video games, or news coverage)

Instructions on how to make or use explosives, unless there is clear context that the content is for a non-violent purpose (for example, clear scientific/educational purpose use or fireworks)

Exposure of vulnerable individuals' identities without their permission

Any content containing statements of intent, calls for action, representation, support or advocating for violence due to voting, voter registration, or the outcome of an election

Misinformation that contributes to imminent violence or physical harm

2. Dangerous Individuals and Organizations

. . . .

We do not allow the following people (living or deceased) or groups to maintain a presence (for example, have an account, Page, Group) on our platform:

Terrorist organizations and terrorists

- A terrorist organization is defined as:
  - Any non-governmental organization that engages in premeditated acts of violence against persons or property to intimidate a civilian population, government, or international organization in order to achieve a political, religious, or ideological aim
- A member of a terrorist organization or any person who commits a terrorist act is considered a terrorist
  - A terrorist act is defined as a premeditated act of violence against persons or property carried out by a non-government actor to intimidate a civilian population, government, or international organization in order to achieve a political, religious, or ideological aim.

Hate organizations and their leaders and prominent members

- A hate organization is defined as:
  - Any association of three or more people that is organized under a name, sign, or symbol and that has an ideology, statements, or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability.

Mass and serial murderers

- We consider a homicide to be a mass murder if it results in four or more deaths in one incident

- We consider any individual who has committed two or more murders over multiple incidents or locations a serial murderer
- We make these assessments based upon the information available to us and will generally apply this policy to a mass or serial murderer who meets any of the following criteria:
  - They were convicted of mass or serial murder.
  - They were killed by law enforcement during commission of the mass or serial murder or during subsequent flight.
  - They killed themselves at the scene or in the aftermath of the mass or serial murder.
  - They were identified by law enforcement with images from the crime.

Human trafficking groups and their leaders

- Human trafficking groups are organizations responsible for any of the following:
  - Prostitution of others, forced/bonded labor, slavery, or the removal of organs
  - Recruiting, transporting, transferring, detaining, providing, harboring, or receiving a minor, or an adult against their will

Criminal organizations and their leaders and prominent members

- A criminal organization is defined as:
  - Any association of three or more people that is united under a name, color(s), hand gesture(s) or recognized indicia, that has engaged in or threatens to engage in criminal activity, including (but not limited to)
    - Homicide
    - Drug trafficking
    - Arms trafficking
    - Identity theft
    - Money laundering
    - Extortion or trafficking
    - Assault
    - Kidnapping
    - Sexual exploitation (covered in section 7 and section 8)

We do not allow symbols that represent any of the above organizations or individuals to be shared on our platform without context that condemns or neutrally discusses the content.

We do not allow content that praises any of the above organizations or individuals or any acts committed by them.

We do not allow coordination of support for any of the above organizations or individuals or any acts committed by them.

3. Promoting or Publicizing Crime

. . . .

Do not post:

Content depicting, admitting, or promoting the following criminal acts committed by you or your associates

- Acts of physical harm committed against people
- Acts of physical harm committed against animals except in cases of hunting, fishing, religious sacrifice, or food preparation/processing
- Poaching or selling endangered species or their parts
- Staged animal vs. animal fights
- Theft
- Vandalism or property damage
- Fraud
- Trafficking as referenced in section 2
- Sexual violence or sexual exploitation, including sexual assault, as referenced in section 7 and section 8

4. Coordinating Harm

. . . .

Do not post:

Statements of intent, calls to action, or advocating for the following:

- Acts of physical harm committed against people
- Acts of physical harm committed against animals except in cases of hunting, fishing, religious sacrifice, or food preparation/processing
- Staged animal vs. animal fights
- Theft
- Vandalism/property damage
- Fraud, defined as the deliberate deception to take advantage of another, secure an unfair gain, or deprive another of money, property, or legal right. Examples of fraud include, but are not limited to:
  - Bribery
  - Embezzlement
  - Money Laundering (concealment of the origins of criminally obtained money)
  - Supporting and/or facilitating the misuse of payment cards
  - Voter fraud, defined as any offers to buy or sell votes with cash or gifts
- Voter suppression, defined as:
  - Misrepresentation of the dates, locations, and times, and methods for voting or voter registration
  - Misrepresentation of who can vote, qualifications for voting, whether a vote will be counted, and what information and/or materials must be provided in order to vote.
  - Other misrepresentations related to voting in an official election may be subject to false news standards, as referenced in section 19
- Arranged marriages with refugees or internally displaced persons
- Trafficking as referenced in section 2
- Sexual violence or sexual exploitation, including sexual assault, as referenced in section 7 and section 8

Offers of services to smuggle or assist in smuggling people.

Content that depicts, promotes, advocates for or encourages participation in a high risk viral challenge, including content with no caption or one that expresses a neutral sentiment.

5. Regulated Goods

. . . .

Do not post:

Content about non-medical drugs (other than alcohol or tobacco) that

- Coordinates or encourages others to sell non-medical drugs
- Depicts, admits to, or promotes sales of non-medical drugs by the poster of the content or their associates
- Promotes, encourages, coordinates, or provides instructions for use of non-medical drugs
- Admits, either in writing or verbally, to personal use of non-medical drugs unless posted in a recovery context

Content that depicts the sale or attempt to purchase marijuana and pharmaceutical drugs. This includes content that

- Mentions or depicts marijuana or pharmaceutical drugs
- Makes an attempt to sell or trade, by which we mean any of the following:
  - Explicitly mentioning the product is for sale or trade or delivery
  - Asking the audience to buy
  - Listing the price
  - Encouraging contact about the product either by explicitly asking to be contacted or including any type of contact information
  - Attempting to solicit the product, defined as:
    - Stating interest in buying the product, or
    - Asking if anyone has the product for sale/trade
- This applies to both individual pieces of content and Pages and Groups primarily dedicated to the sale of marijuana or pharmaceutical drugs

Content that attempts to sell, gift, exchange, or transfer firearms, firearm parts, ammunition, or explosives between private individuals. This includes content that

- Mentions or depicts firearms, firearm parts, ammunition, or explosives and a product unrelated to firearms, and
- Makes an attempt to sell or transfer including any of the following, unless posted by an entity representing a brick-and-mortar store, legitimate website, or brand:
  - Explicitly mentioning the product is for sale or trade or delivery
  - Asking the audience to buy
  - Listing the price or noting that the product is free
  - Encouraging contact about the product either by
    - Explicitly asking to be contacted
    - Including any type of contact information
  - Making an attempt to solicit the item for sale, defined as
    - Stating that they are interested in buying the good, or

- Asking if anyone else has the good for sale/trade

Content that attempts to sell, gift, exchange, transfer, promote or otherwise provide access to instructions for 3D printing or computer-aided manufacturing of firearms or firearm parts.

Content that depicts the trade (buying or selling) of human organs and/or blood where trade is defined as:

- Mentioning or depicting the human organs and/or blood, and
- Indicating that human organs and/or blood are available for selling or buying, or
- Listing a price or expressing willingness to discuss price

Content that encourages contact to facilitate the trade of human organs and/or blood

Content that coordinates or supports the poaching or selling of endangered species and their parts

. . . .

II. Safety

6. Suicide and Self-Injury

. . . .

Do not post:

Content that promotes, encourages, coordinates, or provides instructions for

- Suicide
- Self-injury
- Eating disorders

Content that depicts graphic self-injury imagery

Except in limited situations of newsworthiness, it is against our policies to post content depicting a person who engaged in a suicide attempt or death by suicide

. . . .

7. Child Nudity and Sexual Exploitation of Children

. . . .

Do not post:

Content that depicts participation in or advocates for the sexual exploitation of children, including (but not limited to)

- Engaging in any sexual activity involving minors
- Adults soliciting minors
- Minors soliciting minors
- Minors soliciting adults
- Using our products and site functionality with the intention of sexualizing minors

Content that constitutes or facilitates inappropriate interactions with children, such as

- Initiating unsolicited contact with minors (for example, private messages between stranger adults and minors)
- Soliciting, displaying, sharing, or viewing imagery of nude, sexualized, or sexual activity with minors

- Arranging real-world sexual encounters or obtaining sexual material from a minor directly
- Displaying nudity to minors

Content (including photos, videos, real-world art, digital content, and text) that depicts

- Any sexual activity involving minors
- Minors in a sexual fetish context
- Minors with sexual elements, including (but not limited to):
  - Restraints
  - Focus on genitals
  - Presence of aroused adult
  - Presence of sex toys
  - Sexualized costume
  - Stripping
  - Staged environment (for example, on a bed) or professionally shot (quality/focus/angles)
  - Open-mouth kissing with minor or adult

Content (including photos, videos, real-world art, digital content, and verbal depictions) that shows minors in a sexualized context

Content that depicts child nudity where nudity is defined as

- Visible genitalia (even when covered or obscured by transparent clothing)
- Visible anus and/or fully nude close-up of buttocks
- Uncovered female nipples for children older than toddler-age
- No clothes present from neck to knee for children older than toddler-age
- Digitally-created depictions of nude minors, unless the image is for health or educational purposes

8. Sexual Exploitation of Adults

. . . .

Do not post:

In instances where content consists of any form of non-consensual sexual touching, crushing, necrophilia or bestiality, including:

- Depictions (including real photos/videos), or
- Advocacy (including aspirational and conditional statements), or
- Statements of intent, or
- Calls for action, or
- Participation by yourself or others to engage in any form of the above mentioned sexual acts.

Content that attempts to exploit people by any of the following:

- Coercing money, favors, or images from people by threats of exposure of their naked or semi-naked photos/videos
- Sharing imagery that fulfills all three of the following conditions:
  - Image is non-commercial or produced in a private setting

- Person in the image is (near) nude, engaged in sexual activity, or in a sexual pose
- Lack of consent to share the image is indicated by
  - Vengeful context (for example, caption, comments, or page title)
  - Independent sources (for example, media coverage or law enforcement record)
  - A visible match between the person depicted in the image and the person who has reported the content to us
  - The person who reported the content to us shares the same name as the person depicted in the image

- Sharing imagery of people or a person focusing on sexualized areas of the body such as the breasts, groin, or buttocks (also known as creepshots or upskirts) or focusing on people engaged in sexual activity. The following elements need to be present:
  - the focal point is on a sexualized area of the body or sexual activity, and
  - the person in the image is clearly unaware
- Threatening or stating an intent to share intimate imagery without consent
- Soliciting intimate imagery to view or share without consent
- Threatening or stating an intent to share private sexual conversations

Attempting to coordinate adult commercial sexual services or prostitution activities, such as requesting or offering or asking for rates for escort services and paid sexual fetish or domination services.

. . . .

9. Bullying

. . . .

Do not post:

Content about another private individual that reflects

- Claims about sexual activity
- High-severity physical descriptions
- Ranking individuals on physical appearance or personality
- Threats of non-consensual sexual touching
- Sexualized text targeting another individual
- Attacks through derogatory terms related to sexual activity (e.g. whore, slut)
- An individual in a context that is intended to degrade, for example, menstruating, urinating, vomiting, or defecating
- Physical bullying where the context further degrades the individual
- Comparison to animals that are culturally perceived as intellectually or physically inferior or to an inanimate object

Content that has been photoshopped to target and demean an individual, including by highlighting specific physical characteristics or threatening violence in text or with imagery

Content that specifies an individual as the target of

- Statements of intent to commit violence
- Calls for action of violence
- Statements advocating violence
- Aspirational and conditional statements of violence
- Physical bullying
- Claims about religious identity or blasphemy

In addition, we may remove Pages or Groups that are dedicated to attacking individual(s) by, for example

- Cursing at an individual or individuals
- Making negative character claims
- Making negative ability claims
- Claims about blasphemy
- Appearing to be first person but is actually posted by a different individual than the person referenced and targets more than one individual

We also remove content that is targeted at minors when it contains:

- Cursing
- Claims about romantic involvement or sexual orientation
- Allegations about criminal or illegal behavior
- Coordinating, advocating, or promoting exclusion
- Negative character claims
- Negative ability claims
- Expressions of contempt or disgust
- Calls for death or serious disease or disability
- Videos of physical bullying or violence against minors in a fight context shared with no caption or a neutral or praising caption

In some cases, content is written in the first person but is actually posted by a different individual than the person referenced in the content. This may be done to target the person in the content with the intention of degrading or shaming them. We remove:

Content that contains the following and is reported by the individual depicted:

- Claims about sexual activity
- Comparisons to animals that are culturally perceived as intellectually or physically inferior or to an inanimate object
- High-severity physical descriptions
- Ranking individuals on physical appearance or personality
- Cursing at a person
- Claims about romantic involvement or sexual orientation
- Negative character or ability claims

. . . .

10. Harassment

. . . .

Do not:

Repeatedly contact a single person despite that person's clear desire and action to prevent that contact

Repeatedly contact large numbers of people with no prior solicitation

Send messages to any individual that contain

- Targeted cursing aimed at an individual or group of individuals in the thread
- Calls for death, serious disease or disability, or physical harm aimed at an individual or group of individuals in the thread
- Bullying policy violations
- Claims that a victim of a violent tragedy is lying about being a victim, acting/pretending to be a victim of a verified event, or otherwise is paid or employed to mislead people about their role in the event when sent directly to a survivor and/or immediate family member of a survivor or victim

Send messages to a group that contain any bullying policy violations, regardless of whether the person being targeted is a public or private individual

Target anyone maliciously, including public figures, by

- Attacking them based on their status as a victim of sexual assault or sexual exploitation
- Threatening any participant in public discourse with violence in an attempt to intimidate or silence them
- Calling for self-injury or suicide of a specific person, or group of people
- Attacking them through derogatory terms related to sexual activity (e.g. whore, slut)

Post content about a violent tragedy, or victims of violent tragedies that include claims that a violent tragedy did not occur

Target victims or survivors of violent tragedies by name or by image, with claims that they are

- Lying about being a victim of an event
- Acting/pretending to be a victim of an event
- Otherwise paid or employed to mislead people about their role in the event

Target a minor who is a public figure with:

- Claims about sexual activity or sexually transmitted disease(s)
- Content has been photoshopped to include threats of violence either in text or image (for example, adding bullseye, dart, gun to head)
- Calls for death or serious disease or disability
- Statements of intent to commit violence or low severity harm in an attempt to silence someone
- Objects created to attack through:

- Targeted cursing
- High-severity physical description
- Claims about blasphemy
- Expressions of contempt
- Expressions of disgust

Post content about anyone, including a public figure, that contains a name, photo, or video of the individual and:

- Any statements of intent to commit lethal violence, or
- Any calls for action of lethal violence, or
- Any statements advocating lethal violence

11. Privacy Violations and Impact Privacy Rights

. . . .

Do not post:

Content that facilitates identity theft by posting or soliciting personally identifiable information, including (but not limited to)

- National identification numbers, Social Security numbers, passport numbers, or exam numbers
- Government IDs
- School and education IDs featuring two of the following: (1) name, (2) photo, or (3) ID number
- Digital identities, including passwords

Content that contains medical/psychological, biometric, or genetic records or official documentation of others

Content that facilitates identity theft by sharing personally identifiable information via an external link

Content that facilitates identity theft by sharing private financial information of an organization or business

Content that facilitates identity theft by disclosing the following personal financial information (of either the self or others)

- Bank account and/or card information
- Financial records paired with account information

Content that facilitates identity theft by sharing the private contact information of others defined as

- Private phone numbers or addresses
- Email, Messenger, and chat identities
- The above information may be shared to promote charitable causes, non-violating services, or to facilitate finding missing people or animals

Except in limited cases of newsworthiness, content claimed or confirmed to come from a hacked source, regardless of whether the affected person is a public figure or a private individual.

Content that identifies individuals by name and depicts their personal information, including:

- Driver's licenses, Government IDs other than driver's licenses, Green Cards, or immigration papers
- Marriage, birth, and name change certificates

- Digital identities, including passwords
- License plates

Content that includes photographs that display the external view of private residences if the following conditions apply:

- The residence is a single-family home, or the resident's unit number is identified in the image/caption
- The city or neighborhood is identified
- A resident is mentioned or depicted
- That same resident objects to the exposure of their private residence

Content that exposes the undercover status of law enforcement personnel if

- The content contains the agent's full name or other explicit identification and explicitly mentions their undercover status, or
- The content contains images identifying the faces of the law enforcement personnel and explicitly mentions their undercover status

Content that exposes information about safe houses by sharing any of the below, unless the safe house is actively promoting its location, contact information, or the type of service and protection it offers through comments, posts, Pages or Groups:

- Actual address of the safe house (post box only is allowed),
- Images of the safe house,
- Identifiable city/neighborhood of the safe house, or
- Information outing residents of the safe house

The following content also may be removed

- A reported photo or video of people where the person depicted in the image is:
  - A minor under thirteen years old, and the content was reported by the minor or a parent or legal guardian,
  - A minor between thirteen and eighteen years old, and the content was reported by the minor,
  - An adult, where the content was reported by the adult from outside the United States and applicable law may provide rights to removal
  - Any person who is incapacitated and unable to report the content on their own

III. Objectionable Content

12. Hate Speech

. . . .

Do not post:

Tier 1 attacks, which target a person or group of people who share one of the above-listed characteristics or immigration status (including all subsets except those described as having carried out violent crimes or sexual offenses), where attack is defined as

- Any violent speech or support in written or visual form
- Dehumanizing speech such as reference or comparison to:
  - Insects
  - Animals that are culturally perceived as intellectually or physically inferior
  - Filth, bacteria, disease and feces
  - Sexual predator
  - Subhumanity
  - Violent and sexual criminals
  - Other criminals (including but not limited to "thieves," "bank robbers," or saying "all [protected characteristic or quasi-protected characteristic] are 'criminals' ")
- Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image
- Designated dehumanizing comparisons in both written and visual form

Tier 2 attacks, which target a person or group of people who share any of the above-listed characteristics, where attack is defined as

- Statements of inferiority or an image implying a person's or a group's physical, mental, or moral deficiency
  - Physical (including but not limited to "deformed," "undeveloped," "hideous," "ugly")
  - Mental (including but not limited to "retarded," "cretin," "low IQ," "stupid," "idiot")
  - Moral (including but not limited to "slutty," "fraud," "cheap," "free riders")
- Expressions of contempt or their visual equivalent, including (but not limited to)
  - "I hate"
  - "I don't like"
  - "X are the worst"
- Expressions of disgust or their visual equivalent, including (but not limited to)
  - "Gross"
  - "Vile"
  - "Disgusting"
  - Cursing at a person or group of people who share protected characteristics

Tier 3 attacks, which are calls to exclude or segregate a person or group of people based on the above-listed characteristics. We do allow criticism of immigration policies and arguments for restricting those policies.

Content that describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics.

13. Violence and Graphic Content

. . . .

Do not post:

Imagery of violence committed against real people or animals with comments or captions by the poster that contain

- Enjoyment of suffering
- Enjoyment of humiliation
- Erotic response to suffering
- Remarks that speak positively of the violence; or
- Remarks indicating the poster is sharing footage for sensational viewing pleasure

Videos of dying, wounded, or dead people if they contain

- Dismemberment unless in a medical setting
- Visible internal organs
- Charred or burning people unless in the context of cremation or self-immolation when that action is a form of political speech or newsworthy
- Victims of cannibalism

Videos that show child abuse, defined as

- Repeated kicking, beating, slapping, or stepping on by an adult or animal
- Strangling or suffocating by an adult or animal
- Drowning by an adult or animal
- Biting through skin by an adult or animal
- Poisoning by an adult
- Forcible restraint by an adult
- Inflicting of burn or cut wounds by an adult
- Forcible smoking
- Tossing, rotating, or shaking of an infant (too young to stand) by their wrists/ankles, arms/legs, or neck

. . . .

14. Adult Nudity and Sexual Activity

. . . .

Do not post:

Images of

- Real nude adults, where nudity is defined as
  - ° Visible genitalia except in the context of birth giving and after-birth moments or health-related situations (for example, gender confirmation surgery, genitalia self-examination for cancer or disease prevention/assessment)
  - ° Visible anus and/or fully nude close-ups of buttocks unless photoshopped on a public figure
  - ° Uncovered female nipples except in the context of breastfeeding, birth giving and after-birth moments, health-related situations (for example, post-mastectomy, breast cancer awareness or gender confirmation surgery) or an act of protest

- Sexual activity, including
  - Sexual intercourse
    - Explicit sexual intercourse, defined as mouth or genitals entering or in contact with another person's genitals or anus, where at least one person's genitals are nude
    - Implied sexual intercourse, defined as mouth or genitals entering or in contact with another person's genitals or anus, even when the contact is not directly visible, except in cases of a sexual health context, advertisements, and recognized fictional images or with indicators of fiction
    - Implied stimulation of genitalia/anus, defined as stimulating genitalia/anus or inserting objects into genitalia/anus, even when the activity is not directly visible, except in cases of sexual health context, advertisements, and recognized fictional images or with indicators of fiction
  - Other sexual activities including (but not limited to)
    - Erections
    - Presence of by-products of sexual activity
    - Stimulating genitals or anus, even if above or under clothing
    - Use of sex toys, even if above or under clothing
    - Stimulation of naked human nipples
    - Squeezing naked female breast except in breastfeeding context
  - Fetish content that involves
    - Acts that are likely to lead to the death of a person or animal
    - Dismemberment
    - Cannibalism
    - Feces, urine, spit, snot, menstruation, or vomit

Digital content that meets our definition of sexual activity unless any of the following conditions exist

- Content where the sexual activity (intercourse or other sexual activities) is not directly visible
- Content was posted in a satirical or humorous context
- Content was posted in an educational or scientific context
- Imagery is not sufficiently detailed and only body shapes or contours are visible

15. Sexual Solicitation

. . . .

Do not post:

Attempted coordination of or recruit for adult sexual activities, including but not limited to:

- Filmed sexual activities
- Pornographic activities, strip club shows, live sex performances, erotic dances

- Sexual, erotic, or tantric massages

Explicit sexual solicitation by, including but not limited to the following, offering or asking for:

- Sex or sexual partners
- Sex chat or conversations
- Nude images

Implicit sexual solicitation, which we define as an offer or request to engage paired with at least one of the following elements, none of which violate our Community Standards on their own:

- Suggestive statements, such as "looking for a good time tonight"
- Sexualized slang
- Sexual hints such as mentioning sexual roles, sex positions, fetish scenarios, state of arousal, act of sexual intercourse or activity (sexual penetration or self-pleasuring), commonly sexualized areas of the body such as the breasts, groin, or buttocks, state of hygiene of genitalia or buttocks
- Content (hand drawn or real-world art) that may depict explicit sexual activity or suggestively posed person(s).

An offer or ask for other adult activities such as:

- Commercial pornography
- Nude images, unless modeling context is clear

Sexually explicit language that goes into graphic detail beyond mere reference to:

- A state of sexual arousal (wetness or erection)
- An act of sexual intercourse (sexual penetration, self-pleasuring or exercising fetish scenarios)

16. Cruel and Insensitive

. . . .

Do not post:

Content that depicts real people and mocks their implied or actual serious physical injuries, disease, or disability, non-consensual sexual touching, or premature death

IV. Integrity and Authenticity

17. Spam

. . . .

Do not:

- Artificially increase distribution for financial gain
- Create or use fake accounts or compromise other people's accounts to
  - Impersonate or pretend to be a business, organization, public figure, or private individual
  - Attempt to create connections, create content, or message people
- Restrict access to content by requiring people to like, share, or recommend before viewing

- Encourage likes, shares, or clicks under false pretenses
- Maliciously use login credentials or personally identifiable information by:
  - Attempting to gather or share login credentials or personally identifiable information
  - Using another person's login credentials or personally identifiable information
- Promise non-existent Facebook features

18. Misrepresentation

. . . .

Do not:

Misrepresent your identity by

- Using a name that does not abide by our name policies
- Providing a false date of birth

Misuse our profiles product by

- Creating a profile for someone under thirteen years old
- Maintaining multiple accounts
- Creating inauthentic profiles
- Sharing an account with any other person
- Creating another account after being banned from the site
- Evading the registration requirements outlined in our Terms of Service

Impersonate others by

- Using their images with the explicit aim to deceive people
- Creating a profile assuming the persona of or speaking for another person or entity
- Creating a Page assuming to be or speak for another person or entity for whom the user is not authorized to do so.
- Posting imagery that is likely to deceive the public as to the content's origin, if:
  - The entity or an authorized representative objects to the content, and
  - Can establish a risk of harm to members of the public.

Engage in inauthentic behavior, which includes creating, managing, or otherwise perpetuating

- Accounts that are fake
- Accounts that have fake names
- Accounts that participate in, or claim to engage in, coordinated inauthentic behavior, meaning that multiple accounts are working together to do any of the following:
  - Mislead people about the origin of content
  - Mislead people about the destination of links off our services (for example, providing a display URL that does not match the destination URL)
  - Mislead people in an attempt to encourage shares, likes, or clicks

- ° Mislead people to conceal or enable the violation of other policies under the Community Standards

19. False News

. . . .

20. Memorialization

. . . .

V. Respecting Intellectual Property

21. Intellectual Property

. . . Facebook's Terms of Service do not allow people to post content that violates someone else's intellectual property rights, including copyright and trademark. . . .

VI. Content-Related Requests

22. User Requests

. . . .

23. Additional Protection of Minors

. . . .[231]

231 *Community Standards*, *supra* note 97.

APPENDIX B

Twitter Rules (Synopsis)

Content Boundaries and Use of Twitter

Intellectual property

. . . .

Graphic violence and adult content

[Definitions:]

[1] [G]raphic violence [is] any form of gory media related to death, serious injury, violence, or surgical procedures. . . . [2] [A]dult content [is] any media that is pornographic and/or may be intended to cause sexual arousal. Twitter allows some forms of graphic violence and/or adult content in Tweets marked as containing sensitive media. . . . However, you may not use such content in live video, your profile, or header images. . . . Additionally, Twitter may sometimes require you to remove excessively graphic violence. . . .

Media depicting deceased individuals: We may require you to remove media that depicts the death of an identifiable individual if we receive a request from their family or an authorized representative. . . .

Unlawful use

You may not use our service for any unlawful purposes or in furtherance of illegal activities. . . .

Distribution of hacked materials

We do not permit the use of our services to directly distribute content obtained through hacking that contains personally identifiable information, may put people in imminent harm or danger, or contains trade secrets. . . .

Trends

At times, we may prevent certain content from trending. This includes content that violates the Twitter Rules, as well as content that may attempt to manipulate trends. . . .[232]

These include trends that:

- Contain profanity or adult/graphic references.
- Incite hate on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease.
- Violate the Twitter Rules.

In some cases, we may also consider the newsworthiness of the content, or if it is in the public interest when evaluating potential violations. In these cases, the content might continue to trend on our platform.[233]

Third-party advertising in video content

---

232 *The Twitter Rules*, *supra* note 157.

233 *Twitter Trend FAQs*, TWITTER HELP CTR., https://help.twitter.com/en/using-twitter/twitter-trending-faqs (last visited Mar. 13, 2019).

You may not submit, post, or display any video content on or through our services that includes third-party advertising, such as pre-roll video ads or sponsorship graphics, without our prior consent.

Misuse of Twitter badges
You may not use badges, including but not limited to the "promoted" or "verified" Twitter badges, unless provided by Twitter. . . .

Misuse of usernames
Selling usernames: You may not buy or sell Twitter usernames.
Username squatting: You may not engage in username squatting. . . .

Abusive Behavior
. . . .
Context matters when evaluating for abusive behavior and determining appropriate enforcement actions. Factors we may take into consideration include, but are not limited to whether:

- the behavior is targeted at an individual or group of people;
- the report has been filed by the target of the abuse or a bystander;
- the behavior is newsworthy and in the legitimate public interest.

Violence and physical harm
Violence: You may not[:][234]
[1] [M]ake[ ] violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., "I will kill you".[235]
[2] [W]ish[ ], hope[ ], promote[ ], or express[ ] a desire for death, serious and lasting bodily harm, or serious disease against an entire protected category and/or individuals who may be members of that category. This includes, but is not limited to:

- Hoping that someone dies as a result of a serious disease, e.g., "I hope you get cancer and die."
- Wishing for someone to fall victim to a serious accident, e.g., "I wish that you would get run over by a car next time you run your mouth."
- Saying that a group of individuals deserve serious physical injury, e.g., "If this group of protesters don't shut up, they deserve to be shot."[236]

[3] [T]arget[ ] individuals with content that references forms of violence or violent events where a protected category was the primary target or victims, where the intent is to harass. This includes, but is not limited to sending someone:

- media that depicts victims of the Holocaust;
- media that depicts lynchings.[237]

234 *The Twitter Rules*, *supra* note 157.
235 *Hateful Conduct Policy*, Twitter Help Ctr., https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (last visited, Mar. 13, 2019).
236 *Id.*
237 *Id.*

[4] [A]ffiliate with organizations that—whether by their own statements or activity both on and off the platform—use or promote violence against civilians to further their causes.
Suicide or self-harm: You may not promote or encourage suicide or self-harm. . . .
Child sexual exploitation: You may not promote child sexual exploitation. . . .

Abuse and hateful content
Abuse: You may not engage in the targeted harassment of someone, or incite other people to do so. . . . [This includes] attempt[s] to harass, intimidate, or silence someone else's voice.
Unwanted sexual advances: You may not direct abuse at someone by sending unwanted sexual content, objectify[ ] [someone] in a sexually explicit manner, or otherwise engaging in sexual misconduct.
Hateful conduct: [1] You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. . . .[238]
[2] [You may not] target[ ] individuals with content intended to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., "all [religious group] are terrorists".[239]
[3] [You may not] target[ ] individuals with repeated slurs, tropes or other content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes about a protected category. This includes targeted misgendering or deadnaming of transgender individuals.[240]
Hateful imagery[241] and display names: [1] You may not use hateful images or symbols in your profile image, profile header. [2] You may not . . . use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.[242]

---

238 *The Twitter Rules*, *supra* note 157.
239 *Hateful Conduct Policy*, *supra* note 235.
240 *Id.*
241 Twitter's hateful conduct policy defines hateful imagery as

> logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, gender identity or ethnicity/national origin. Some examples of hateful imagery include, but are not limited to:
>
> - symbols historically associated with hate groups, e.g., the Nazi swastika;
> - images depicting others as less than human, or altered to include hateful symbols, e.g., altering images of individuals to include animalistic features; or
> - images altered to include hateful symbols or references to a mass murder that targeted a protected category, e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust.

Hateful Conduct Policy, *supra* note 235.
242 *The Twitter Rules*, *supra* note 157.

[3] [You may not] send[ ] an individual unsolicited hateful imagery . . . .[243]

Private information and intimate media

Private information: You may not publish or post other people's private information without their express authorization and permission. . . .

Intimate media: You may not post or share intimate photos or videos of someone that were produced or distributed without their consent. . . . Note: limited exceptions may apply . . . .

Threats to expose / hack: [1] You may not threaten to expose someone's private information or intimate media. [2] You . . . may not threaten to hack or break into someone's digital information or attempt to incentivize others to do so (e.g., through setting a bounty or reward on such actions).

Impersonation

[1] You may not impersonate individuals, groups, or organizations in a manner that is intended to or does mislead, confuse, or deceive others. . . . [2] [Y]ou may [not] maintain parody, fan, commentary, or newsfeed accounts . . . if the intent of the account is to engage in spamming or abusive behavior.[244]

243 *Hateful Conduct Policy*, *supra* note 235.

244 *The Twitter Rules*, *supra* note 157.