The Varieties of Counterspeech and Censorship on Social Media

Dawn Carla Nunziato[*]

TABLE OF CONTENTS

The year 2020 was without a doubt a remarkable and unprecedented one, on many accounts and for many reasons. Among other reasons, it was a year in which the major social media platforms extensively experimented with the adoption of a variety of new tools and practices to address grave problems resulting from harmful speech on their

 William Wallace Kirkpatrick Research Professor and Professor of Law, The George Washington University Law School. I am grateful to the participants in this symposium for their comments on this Article, as well as to editors at the UC Davis Law Review for their excellent editorial assistance, and to Chris Frascella and Lucy Xiong for excellent and expert research assistance in connection with this Article. I am also grateful to Dean Dayna Matthew for financial support of my research and writing.

platforms — notably, the vast amounts of misinformation associated with the COVID-19 pandemic and with the 2020 presidential election and its aftermath. By and large — consistent with First Amendment values of combatting bad speech with good speech — the platforms sought to respond to harmful online speech by resorting to different types of flagging, fact-checking, labeling, and other forms of counterspeech. Only when confronting the most egregiously harmful types of speech did the major platforms implement policies of censorship or removal — or the most extreme response of deplatforming speakers entirely. In this Article, I examine the major social media platforms' experimentation with a variety of approaches to address the problems of political and election-related misinformation on their platforms — and the extent to which these approaches are consistent with First Amendment values. In particular, I examine what the major social media platforms have done and are doing to facilitate, develop, and enhance counterspeech mechanisms on their platforms in the context of major elections, how closely these efforts align with First Amendment values, and measures that the platforms are taking, and should be taking, to combat the problems posed by filter bubbles in the context of the microtargeting of political advertisements.

This Article begins with an overview of the marketplace of ideas theory of First Amendment jurisprudence and the crucial role played by counterspeech within that theory. I then analyze the variety of types of counterspeech on social media platforms — by users and by the platforms themselves — with a special focus on the platforms' counterspeech policies on elections, political speech, and misinformation in political/campaign speech specifically. I examine in particular on the platforms' prioritization of labeling, fact-checking, and referring users to authoritative sources over the use of censorship, removal, and deplatforming (which the platforms tend to reserve for the most harmful speech in the political sphere and which they ultimately wielded in the extraordinary context of the speech surrounding the January 2021 insurrection). I also examine the efforts that certain platforms have taken to address issues surrounding the microtargeting of political advertising, issues which are exacerbated by the filter bubbles made possible by segmentation and fractionation of audiences in social media platforms.

I. THE MARKETPLACE OF IDEAS AND THE ROLE OF COUNTERSPEECH IN OUR FIRST AMENDMENT JURISPRUDENCE

The "marketplace of ideas" or "free trade in ideas" model has long been acknowledged as the preeminent model on which our First

Amendment free speech protections are based. Although courts sometimes credit other justifications for protecting speech — including the role of free speech in our system of democratic self-government[1] and in advancing individuals' interest in self-expression and self-fulfillment[2] — the courts' preeminent and most frequently invoked justification or model for freedom of expression is the marketplace of ideas. The very notion of ideas vying and competing in the market presupposes an interplay and exchange of ideas and therefore presupposes that "counterspeech" will be made available in response to speech and that the citizenry will be able to hear and receive competing viewpoints and perspectives. Accordingly, pursuant to this predominant model, the default response to bad speech is not censorship but more (better) speech. As Justice Brandeis explained in his oft-quoted concurrence in *Whitney v. California*: "If 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."[3] Under the marketplace theory of the First Amendment, the default remedy for harmful ideas in the marketplace of speech is not censorship, but counterspeech, which recognizes the importance of access to diverse, antagonistic, competing viewpoints and the free trade in ideas, which functions by allowing those who are exposed to bad speech to be exposed to good speech as a counterweight. The counterspeech mechanism provides that this default remedy applies broadly in most circumstances, except in the case of "emergency," i.e., in circumstances where there is not sufficient time "to avert the evil by process of education." Thus, we see *Brandenburg v. Ohio*'s formulation allowing for speech constituting "incitement" to be restricted because of the likelihood that it will lead to imminent harm.[4]

Notably, while the marketplace of ideas theory (and its default response of counterspeech as a remedy for bad speech) accords broad protection to good and bad *ideas*, this theory does not accord the same broad protections to good and bad assertions of *fact*[5] (such as assertions

[1] *See* Thomas I. Emerson, *Toward a General Theory of the First Amendment*, 72 YALE L.J. 877, 878 (1963).

[2] As David Richards explains, the First Amendment rests not only on the value of creating an informed electorate, but also "rests . . . on the deeper moral premises regarding the general exercise of autonomous expressive and judgmental capacity and the good that this affords in human life." David. A. J. Richards, *Free Speech and Obscenity Law: Toward a Moral Theory of the First Amendment*, 123 U. PA. L. REV. 45, 68 (1974).

[3] Whitney v. California, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring).

[4] *See* Brandenburg v. Ohio, 395 U.S. 444, 447 (1969).

[5] Dawn Carla Nunziato, *The Marketplace of Ideas Online*, 94 NOTRE DAME L. REV. 1519, 1526 (2019) [hereinafter *The Marketplace of Ideas Online*].

that one can vote by text or that the election will be held on Wednesday instead of Tuesday). The Supreme Court, in embracing the marketplace of ideas theory, has made clear that there is no such thing as a false *idea* — that all *ideas* are protected — but it has also explained that false statements of *fact* are not similarly protected.[6] While the Court has sometimes recognized the minimal potential contributions to the marketplace of ideas made by harmless lies[7] and by some incidental/inevitable false statements of fact,[8] it has also emphasized that the First Amendment does not stand in the way of regulating intentionally false or misleading assertions of fact.[9] In sum, the predominant marketplace of ideas theory of the First Amendment accords broad protection to ideas — even bad ones — and provides that counterspeech, not censorship, is the default response to harmful speech consisting of harmful *ideas* — but this theory does not accord similarly broad protection to all *assertions of fact*.

The Supreme Court has specifically recognized the importance of counterspeech as an integral element of our First Amendment marketplace of ideas model in a number of cases throughout the past century, even as the mediums for expression have shifted from print to broadcast to cable to the Internet. In addition, and relatedly, the Court has recognized the importance of citizens being exposed to and confronting a diverse array of opinions — including speech with which they disagree. Although the enablement and prevalence of filter bubbles on the Internet is a recent development, First Amendment

[6] Gertz v. Robert Welch, Inc., 418 U.S. 323, 339-40 (1974); Nunziato, *The Marketplace of Ideas Online*, *supra* note 5, at 1526.

[7] *See* United States v. Alvarez, 567 U.S. 709, 732 (2012) (Breyer, J., concurring). In *United States v. Alvarez*, the Supreme Court, in a 6-3 decision, struck down a portion of the Stolen Valor Act, a federal law that criminalized the making of false statements about having a military medal. The Act made it a misdemeanor to falsely represent oneself as having received any U.S. military decoration or medal and provided for prison terms up to six months (and up to one year if the subject of such lies was the Medal of Honor). In a challenge brought by Xavier Alvarez, who was convicted under the Act for publicly lying about receiving the Congressional Medal of Honor, the Court struck down the Stolen Valor Act on First Amendment grounds. Justice Kennedy, writing for a plurality, held that harmless false statements are not, by the sole reason of their falsity, excluded from First Amendment protection. *See also id.* at 711 (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 remedied by counterspeech — if a list of medal recipients were made available on the Internet).

[8] *See* N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279 (1964).

[9] *See Gertz*, 418 U.S. at 340 ("[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." (citations omitted)).

jurisprudence has long been centered around the importance of citizens' exposure to diverse, antagonistic, and competing viewpoints.

In its decisions constitutionalizing the common law of defamation, for example, the Court has recognized our "profound national commitment to the principle that debate on public issues should be uninhibited, robust, and wide open"[10] and accordingly that the default response to speech critical of public figures is not liability or retraction/removal, but is instead more speech. As the Court explained in *Gertz v. Welch*, "the first remedy available [to targets of defamation] is self-help – using available opportunities to contradict the lie or correct the error."[11] This remedy is especially suited to public officials and public figures, who generally "enjoy significantly greater access to the channels of effective communication and hence have . . . a realistic opportunity to counteract false statements [and engage in counterspeech]."[12]

The Court has further recognized the importance of counterspeech and broad exposure to competing viewpoints in upholding the "Fairness Doctrine," a set of regulations that required broadcast television and radio stations to provide fair coverage to competing sides of the discussion of public issues and to provide for a "reasonable opportunity to respond" if "an attack is made on the honesty, character, integrity, or like personal goals of an identified person or group."[13] In upholding the constitutionality of the Fairness Doctrine's obligations placed upon broadcasters, the Court emphasized the First Amendment goal of "producing an informed public capable of conducting its own affairs" and refused to allow forums in broadcast television or radio to be turned into information silos monopolized by one side of the debate or the other on controversial issues of public importance.[14] Instead, the Court recognized the importance of facilitating opportunities for speech and counterspeech in order for our information ecosystem to produce "an informed public capable of conducting its own affairs."[15]

Similarly, in *Turner Broadcasting System v. FCC*,[16] the Court recognized the importance of ensuring that citizens are exposed to competing and diverse viewpoints. *Turner* involved a challenge brought by several cable systems operators to the "must carry" provisions of the

[10] *Sullivan*, 376 U.S. at 270.

[11] *Gertz*, 418 U.S. at 344.

[12] *Id.*

[13] Red Lion Broad. Co. v. FCC, 395 U.S. 367, 373-74 (1969).

[14] *Id.* at 392.

[15] *Id.*

[16] Turner Broad. Sys. v. FCC, 512 U.S. 622 (1994).

Cable Television Consumer Protection and Competition Act of 1992 (the Cable Act).[17] The Act required cable systems operators to carry the signals of local educational public broadcast television stations, without charge, in the same numerical channel position as when these programs were broadcast over the air.[18] The Court credited several important government interests that were advanced by the Act, including a government purpose "of the highest order in ensuring public access to a multiplicity of information sources."[19] On this point, the Court explained that "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."[20] Upholding the Act on remand, Justice Breyer credited the Act's purpose of advancing the national communications policy of protecting "the widest possible dissemination of information from diverse and antagonistic sources" and explained that:

> [This national communications] policy, in turn, seeks to facilitate the public discussion and informed deliberation, which, as Justice Brandeis pointed out many years ago, democratic government presupposes and the First Amendment seeks to achieve. . . . Indeed, *Turner* [below] rested in part upon the proposition that "assuring that the public has access to a multiplicity of information sources is a governmental purpose of the highest order, for it promotes values central to the First Amendment."[21]

In its public forum jurisprudence as well, the Court has emphasized the importance of citizens' exposure to competing viewpoints, including diverse, conflicting, and antagonistic ones. In recognizing the important role that public forums like streets and sidewalks serve in our First Amendment jurisprudence, the Court recently explained that:

> It is no accident that public streets and sidewalks have developed as venues for the exchange of ideas. Even today, they remain one of the few places where a speaker can be confident that he is not simply preaching to the choir. With respect to other means of communication, an individual confronted with

[17] *Id.* at 632-34.

[18] *See id.* at 630-31.

[19] *Id.* at 663.

[20] *Id.*

[21] Turner Broad. Sys. v. FCC, 520 U.S. 180, 226-27 (1997) (Breyer, J., concurring) (quoting *Turner*, 512 U.S. at 663).

> an uncomfortable message can always turn the page, change the channel, or leave the Web site. Not so on public streets and sidewalks. There, a listener often encounters speech he might otherwise tune out. In light of the First Amendment's purpose "to preserve an uninhibited marketplace of ideas in which truth will ultimately prevail," this aspect of traditional public fora is a virtue, not a vice.[22]

In addition, in the handful of cases in which the Supreme Court has analyzed the marketplace of ideas and specifically in the Internet context, it has further recognized the importance of counterspeech and exposure to diverse viewpoints to our information ecosystem. In *Packingham v. North Carolina*, for example, Justice Kennedy emphasized the important role served by social media platforms in today's marketplace of ideas, explaining that they serve as modern day public forums in which citizens can access, engage with, and challenge their elected representatives.[23] Justice Kennedy specifically identified Facebook and Twitter as serving these roles, explaining that Twitter in particular is a forum where "users can petition their elected representatives and otherwise engage with them in a direct manner," as "Governors of all 50 States and almost every Member of Congress" utilize Twitter as a forum in which to engage their constituents.[24] Kennedy further noted that social media sites offer "relatively unlimited, low-cost capacity for communication of all kinds," where users can "engage in a wide array of protected First Amendment activity on topics 'as diverse as human thought.'"[25] He observed that the Internet in general and social media in particular are "integral to the fabric of modern society and culture"[26] as they constitute "principal sources for . . . speaking and listening in the modern public square, . . . [and are]

---

[22] McCullen v. Coakley, 573 U.S. 464, 476 (2014) (citation omitted) (quoting FCC v. League of Women Voters of Cal., 468 U.S. 364, 377 (1984)). On this point, see also Cass Sunstein, who explains that one of the most important functions of the public forum doctrine is to ensure the opportunity for "shared exposure to diverse speakers with diverse views and complaints." CASS R. SUNSTEIN, #REPUBLIC: DIVIDED DEMOCRACY IN THE AGE OF SOCIAL MEDIA 38 (2018). Sunstein explains that in order for us to meet the demands of citizenship in a deliberative democracy, we must be exposed to a diverse set of topics and opinions. *Id.*

[23] *See* Packingham v. North Carolina, 137 S. Ct. 1730, 1735 (2017).

[24] *Id.*

[25] *Id.* at 1735-36 (quoting Reno v. ACLU, 521 U.S. 844, 870 (1997)).

[26] *Id.* at 1738.

perhaps the most powerful mechanisms available to a private citizen to make his or her voice heard."[27]

Indeed, courts have continued to recognize the importance of social media platforms as public forums for discussions with public officials of matters of public importance. In the Trump Twitter blocking case,[28] the Second Circuit emphasized in particular the role of social media platforms in serving as forums for counterspeech on matters of public importance. Accordingly, the court refused to allow President Trump to use his Twitter platform to create a forum only containing speech that is favorable to him. Trump famously sought to use his Twitter platform to allow only followers whose comments were favorable to him and his policies to follow him, while blocking those who criticized or disagreed with him. Trump made use of his Twitter platform to share news and announcements of public importance (like the hiring and firing of officials, the announcement of executive orders, etc.), but sought to restrict who could follow this account. For example, Trump used his account to announce his intention to nominate Christopher Wray for the position of FBI director,[29] as well as to remove then-Secretary of State Rex Tillerson,[30] then-Secretary of Veterans Affairs David Shulkin,[31] then-Secretary of Defense Mark Esper,[32] and then-Director of the Cybersecurity and Infrastructure Security Agency ("CISA") Chris

[27] *Id.* at 1737.

[28] Knight First Amend. Inst. at Columbia Univ. v. Trump, 928 F.3d 226 (2d Cir. 2019), *vacated and dismissed as moot sub nom.* Biden v. Knight First Amend. Inst. at Columbia Univ., 141 S. Ct. 1220 (2021).

[29] Donald J. Trump (@realDonaldTrump), TWITTER (June 7, 2017, 4:44 AM), https://twitter.com/realdonaldtrump/status/872419018799550464 [https://perma.cc/JWH6-9EZR] ("I will be nominating Christopher A. Wray, a man of impeccable credentials, to be the new Director of the FBI. Details to follow.").

[30] Donald J. Trump (@realDonaldTrump), TWITTER (Mar. 13, 2018, 5:44 AM), https://twitter.com/realDonaldTrump/status/973540316656623616 [https://perma.cc/8TM5-XCTQ] ("Mike Pompeo, Director of the CIA, will become our new Secretary of State. He will do a fantastic job! Thank you to Rex Tillerson for his service!").

[31] Donald J. Trump (@realDonaldTrump), TWITTER (Mar. 28, 2018, 2:31 PM), https://twitter.com/realDonaldTrump/status/979108846408003584 [https://perma.cc/73UA-B9BB] ("I am pleased to announce that I intend to nominate highly respected Admiral Ronny L. Jackson, MD, as the new Secretary of Veterans Affairs . . ." *immediately followed by* ". . . In the interim, Hon. Robert Wilkie of DOD will serve as Acting Secretary. I am thankful to Dr. Shulkin's service to our country and to our GREAT VETERANS!").

[32] Donald J. Trump (@realDonaldTrump), TWITTER (Nov. 9, 2020, 12:54 PM), https://twitter.com/realDonaldTrump/status/1325859407620689922 [https://archive.is/mb0Y2] (". . . Chris will do a GREAT job! Mark Esper has been terminated. I would like to thank him for his service.").

Krebs[33] from their respective positions, and to announce that the United States Government would no longer accept or allow transgender individuals to serve in the military.[34] Yet, Trump sought to allow only favorable followers and favorable Twitter commentary on such announcements and decisions — and to block the counterspeech of those who disagreed with him and his policies. In defending such actions against constitutional attack, he claimed that his Twitter account was private, not governmental, or in the alternative, that the interactive comment spaces associated with his tweets constituted "government speech" immune from scrutiny under the Free Speech Clause of the First Amendment — instead of a public forum in which the citizenry was permitted to counter his speech with criticisms and comments and diverse and antagonistic viewpoints.

The Second Circuit disagreed with Trump, and held instead that the "interactive space" associated with the president's tweets constituted a public forum because it was a forum "in which other users may directly interact with the content of the tweets by . . . replying to, retweeting or liking the tweet."[35] In holding that the President, by speaking on Twitter, created a public forum consisting of this interactive space, the court concluded that Trump's act of blocking users from speaking in this space amounted to unconstitutional viewpoint discrimination within a public forum.[36] Referencing the important role played by social

[33] Donald J. Trump (@realDonaldTrump), TWITTER (Nov. 17, 2020, 7:07 PM), https://twitter.com/realDonaldTrump/status/1328852354049957888 [https://archive.is/1gN5x] (". . . votes from Trump to Biden, late voting, and many more. Therefore, effective immediately, Chris Krebs has been terminated as Director of the Cybersecurity and Infrastructure Security Agency.").

[34] Donald J. Trump (@realDonaldTrump), TWITTER (July 26, 2017, 5:55 AM), https://twitter.com/realDonaldTrump/status/890193981585444864 [https://perma.cc/7WRL-TSM6] ("After consultation with my Generals and military experts, please be advised that the United States Government will not accept or allow . . ."); Donald J. Trump (@realDonaldTrump), TWITTER (July 26, 2017, 6:04 AM), https://twitter.com/realDonaldTrump/status/890196164313833472 [https://perma.cc/2KA8-ZTAR] (". . . Transgender individuals to serve in any capacity in the U.S. Military.").

[35] *See* Knight First Amend. Inst. at Columbia Univ. v. Trump, 928 F.3d 226, 233 (2d Cir. 2019) (quoting Knight First Amend. Inst. at Columbia Univ. v. Trump, 302 F. Supp. 3d 541, 579 (S.D.N.Y. 2018)).

[36] *See id.* at 233-34. Of special note is Justice Thomas's concurrence in the Supreme Court's decision vacating the Second Circuit's decision in this case as moot, in which Thomas expresses concern not about former President Trump's ability to restrict speech on the basis of viewpoint in connection with his Twitter account, but with Twitter's far vaster ability to restrict President Trump's speech on the Twitter platform — by deplatforming the president and therefore permanently banning him from communicating with eighty-nine million followers on this platform. As Justice Thomas observes:

media companies like Twitter in facilitating counterspeech in the online marketplace of ideas, the court observed:

> [W]e write at a time in the history of this nation when the conduct of our government and its officials is subject to wide-open, robust debate. This debate encompasses an extraordinarily broad range of ideas and viewpoints and generates a level of passion and intensity the likes of which have rarely been seen. This debate, as uncomfortable and as unpleasant as it frequently may be, is nonetheless a good thing. In resolving this appeal, we remind the litigants and the public that *if the First Amendment means anything, it means that the best response to disfavored speech on matters of public concern is more speech, not less.*[37]

In short, as forums for speech have evolved in the past century from print to broadcast to the Internet and social media, the courts have continued to recognize the preeminent importance of the marketplace of ideas, of broad exposure to diverse, competing, and antagonistic viewpoints, and specifically of counterspeech as essential to "producing

> The disparity between Twitter's control and Mr. Trump's control [of speech on the Twitter platform] is stark, to say the least. Mr. Trump blocked several people from interacting with his messages. Twitter barred Mr. Trump not only from interacting with a few users, but removed him from the entire platform, thus barring all Twitter users from interacting with his messages. Under its terms of service, Twitter can remove any person from the platform — including the President of the United States — "at any time for any or no reason." . . . Today's digital platforms provide avenues for historically unprecedented amounts of speech, including speech by government actors. Also unprecedented, however, is the concentrated control of so much speech in the hands of a few private parties. . . . The Second Circuit feared that then-President Trump cut off speech by using the features that Twitter made available to him. But if the aim is to ensure that speech is not smothered, then the more glaring concern must perforce be the dominant digital platforms themselves. As Twitter made clear, the right to cut off speech lies most powerfully in the hands of private digital platforms. The extent to which that power matters for purposes of the First Amendment and the extent to which that power could lawfully be modified raise interesting and important questions.
>
> Biden v. Knight First Amend. Inst. at Columbia Univ., 141 S. Ct. 1220, 1221-27 (2021) (Thomas, J., concurring) (internal citations and footnotes omitted).

[37] *Knight First Amend. Inst.*, 928 F.3d at 240 (emphasis added).

an informed public capable of conducting its own affairs," and that "the best response to disfavored speech . . . is more speech, not less."[38]

II. THE VARIETIES OF COUNTERSPEECH FACILITATED ON THE INTERNET

In today's information ecosystem, the social media platforms facilitate speech and counterspeech on the Internet on an unprecedented scale. The platforms do this, first and foremost, by serving as forums where individuals can speak and respond to one another's speech. This occurs notably on social media platforms like Twitter and Facebook, where users can respond to one another's speech and where the default remedy for bad speech, by and large, continues to be counterspeech not censorship. Although these platforms host a vast array of speech — including a vast array of harmful speech — by and large the platforms' primary response to harmful speech has not been censorship/removal but rather counterspeech of one form or another. While the major platforms do censor some limited categories of harmful speech — and have done so to a much greater extent in the context of the January 6, 2021, insurrection at the Capitol — they have generally adhered to First Amendment values of allowing harmful speech, absent such a likelihood of imminent harm or emergency, and facilitating counterspeech as the default response to harmful speech. Because the major social media platforms wield enormous power over speech in today's information ecosystem, those who control such platforms have expressed reticence to exercise the powers of the censor and have sought to manifest their commitment to the free speech and counterspeech ideals on which the First Amendment is premised. For many years, for example, Twitter characterized itself as "the free speech wing of the free speech party"[39] and rejected the role of being the arbiter of truth in the online information ecosystem, deferring instead to its role of facilitating the marketplace of ideas and counterspeech as a remedy for bad speech on its platform.[40] And Facebook, for its part, has

---

[38] *See* Red Lion Broad. Co. v. FCC, 395 U.S. 367, 392 (1969); *Knight First Amend. Inst.*, 928 F.3d at 240.

[39] Marvin Ammori, *The "New"* New York Times*: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2260 (2014); Josh Halliday, *Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party*,*'* GUARDIAN (Mar. 22, 2012), https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech [https://perma.cc/HV26-QF4H].

[40] *See* Colin Crowell, *Our Approach to Bots and Misinformation*, TWITTER BLOG (June 14, 2017), https://blog.twitter.com/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html [https://perma.cc/K82F-HT3Z] ("Twitter's open and real-time

publicly railed against the idea that it should be the arbiter of truth on matters of public importance, explaining: "in a democracy, people should decide what is credible, not tech companies"[41] and "[w]e don't believe . . . that it's an appropriate role for us to referee political debates."[42] Accordingly, the major platforms generally have sought to limit their censoring/removal interventions to the most harmful speech — such as speech that incites immediate violence or harm, constitutes an actual threat of violence,[43] or contains child sex abuse materials — categories of speech that are likewise unprotected in First Amendment jurisprudence.[44] In addition, the platforms have been more willing to wield their censoring/removal interventions in the context of harmful and false assertions of fact (e.g., false information about how or when to vote, medical misinformation surrounding the COVID-19 pandemic). This too is consistent with First Amendment jurisprudence and the marketplace of ideas model, which broadly extends protections to good and bad *ideas*, but less so to good and bad assertions of *fact*, as discussed above. Yet overwhelmingly, the platforms' predominant and preferred response to the vast array of "bad" speech on their platforms had been through the mechanism of counterspeech. This is particularly true regarding the online speech by public officials, including President Trump, which the platforms had been quite reticent to outright censor until the events surrounding Trump's incendiary speech inciting the insurrection at the Capitol on January 6, 2021. Prior to the insurrection,

nature is a powerful antidote to the spreading of all types of false information. This is important because we cannot distinguish whether every single Tweet from every person is truthful or not. We, as a company, should not be the arbiter of truth. [Instead, we look to] journalists, experts and engaged citizens [who] Tweet side-by-side correcting and challenging public discourse in seconds.").

[41] *See* David Klepper, *Facebook Clarifies Zuckerberg Remarks on False Political Ads*, AP NEWS (Oct. 24, 2019), https://apnews.com/64fe06acd28145f5913d6f815bec36a2 [https://perma.cc/4GZQ-K37D].

[42] *See* Nick Clegg, *Facebook, Elections and Political Speech*, FACEBOOK NEWSROOM (Sept. 24, 2019), https://about.fb.com/news/2019/09/elections-and-political-speech/ [https://perma.cc/Z9VZ-Q6FU] [hereinafter *Facebook, Elections and Political Speech*]. Facebook will, however, subject the posts of political action committees and political advocacy groups to its fact-checking process. Facebook has explained that while it will not fact-check political ads from candidates, it does evaluate the accuracy of political ads from political advocacy groups or political action committees. *See* Klepper, *supra* note 41.

[43] *See Violent Threats Policy*, TWITTER (Mar. 2019), https://help.twitter.com/en/rules-and-policies/violent-threats-glorification [https://perma.cc/2V9L-27KB] (discussing Twitter's policy on violent threats).

[44] In contrast to First Amendment jurisprudence, however, the major social media platforms generally have taken a much stricter position on restricting hate speech than under First Amendment law.

the platforms had been much more inclined to engage in counterspeech in response to harmful speech by public officials. Along these lines, Twitter had created and implemented a "public interest" exception[45] to its otherwise applicable rules that would require removal of certain categories of harmful speech for the tweets of elected and government officials, an exception that it had justified based on the "significant public interest in knowing and being able to discuss [elected and government officials'] actions and statements."[46] The platforms' former policies embodying a reticence to censor/remove the speech of public officials (and more broadly, speech in the public interest) was generally consistent with First Amendment values, in which speech of public officials and on matters of public importance is within the core of the First Amendment's protections.

The counterspeech facilitated by the major social media platforms comes in a variety of forms. First and foremost, the platforms facilitate counterspeech by creating and hosting forums for platforms where people can respond directly or indirectly to one another's speech. Twitter and Facebook, for example, create vast opportunities for counterspeech in the forms of people interacting with one another's speech by following or friending one another and interacting with one another's speech in a wide variety of ways, such as commenting in response, retweeting, liking/disliking, etc. While private figure users can control who they follow and who follows them by blocking, muting, unfriending and the like (which in turn can lead to problems for the counterspeech mechanism caused by "filter bubbles"), such control does not apply to public officials, whose use of social media platforms like Facebook or Twitter has been held by courts to create public forum in which followers cannot be blocked on the basis of their viewpoint.[47]

---

[45] *About Public-Interest Exceptions on Twitter*, TWITTER, https://help.twitter.com/en/rules-and-policies/public-interest (last visited Jan. 25, 2021) [https://perma.cc/KG7Y-QBW4].

[46] *Id.* For example, even though President Trump's infamous tweet "When the looting starts, the shooting starts" violated Twitter's policy prohibiting the glorification of violence, the tweet was hidden behind a notice claiming it breached Twitter's policies on glorifying violence. The tweet could still be viewed and retweeted with the comment, but could not be liked, replied to, or retweeted otherwise. *See* Ryan Browne, *Twitter Flags Trump Tweet on Minneapolis for 'Glorifying Violence*,' CNBC (May 29, 2020, 3:46 AM EST), https://www.cnbc.com/2020/05/29/twitter-flags-trump-tweet-on-minneapolis-for-glorifying-violence.html [https://perma.cc/7EJD-KH38].

[47] *See, e.g.*, Knight First Amend. Inst. at Columbia Univ. v. Trump, 928 F.3d 226 (2d Cir. 2019) (holding that President Trump's use of the interactive features of his @realDonaldTrump Twitter account for government purposes created a limited public forum in which viewpoint discrimination was prohibited and therefore Trump's act of blocking Twitter followers who were critical of him violated the First Amendment);

This is one area in which First Amendment law has been successfully invoked to prevent the creation of information silos or filter bubbles.

Second, the platforms engage in and facilitate counterspeech themselves, including by (1) labeling speech (with labels determined by the platforms themselves and/or by working with external fact-checkers to determine and engage in such labeling); (2) providing and directing users to authoritative third-party information/external trusted sources in response to speech; (3) referring speech to external fact-checkers for evaluation and subsequent labeling; and (4) referring speech to external fact-checkers/external trusted sources for evaluation and commissioning the production of responsive counterspeech (such as Facebook's "related articles," discussed below). Such counterspeech interventions by the platforms have developed extensively in the past two years and expanded dramatically in the months leading up to the 2020 presidential election. Below I analyze in detail these interventions in the context of political and election-related speech surrounding the 2020 presidential elections, then turn to an analysis of the efficacy of these counterspeech interventions by the platforms.

III. ELECTION AND POLITICAL SPEECH, "COUNTERSPEECH" RESPONSES, AND CENSORSHIP OF SUCH SPEECH BY THE MAJOR PLATFORMS

In the time period leading up to and immediately after the 2020 presidential election, the major platforms undertook extensive measures to check, counter, and, in some extreme circumstances, to remove election-related misinformation and disinformation. First, the platforms adopted policies regarding misinformation that they applied to political and campaign related speech, including policies applicable to manipulated media such as deepfakes and shallow fakes. In addition, the platforms enacted extensive policies regarding misinformation about the logistics of the voting process and the post-Election Day announcement of results. The platforms generally wielded their power consistent with the approach described above, by censoring/removing only the most egregious and harmful false posts, while engaging in

Davison v. Randall, 912 F.3d 666 (4th Cir. 2019) (holding that the Chair of the Board of Supervisors for Loudoun County, Virginia, Phyllis Randall, created a limited public forum by using her Facebook page for government purposes and violated the First Amendment by engaging in unconstitutional viewpoint discrimination when she blocked a constituent from posting on her page because his comments were critical of her). *See generally* Dawn Carla Nunziato, *From Town Square to Twittersphere: The Public Forum Doctrine Goes Digital*, 25 B.U. J. SCI. & TECH. L. 1, 43-54 (2019) (discussing and summarizing cases in which government officials' acts of blocking constituents on social media sites was at issue).

various forms of counterspeech with respect to posts deemed less harmful and with respect to posts by government officials on matters of public importance. The platforms adopted this approach until the unprecedented actions of former President Trump and his surrogates in the wake of the 2020 election and the events surrounding the January 2021 insurrection, as I examine below. Then, the platforms modified their deferential approach of generally exempting posts by government officials from content moderation measures in response to the unprecedented false and harmful information about the results of the election posted by former President Trump and his surrogates and in response to Trump's posts inciting the insurrection at the Capitol.

A. Twitter

With respect to manipulated media like deepfakes and shallow fakes,[48] Twitter generally takes the approach of prioritizing counterspeech or labeling instead of censorship or removal. In February 2020, Twitter adopted a policy on "synthetic and manipulated media" that provides: "You may not deceptively share synthetic or manipulated media that are likely to cause harm"[49] and explained: "we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context."[50] Pursuant to this policy, and consistent with First Amendment values

[48] *See* Yoel Roth & Ashita Achuthan, *Building Rules in Public: Our Approach to Synthetic & Manipulated Media*, TWITTER BLOG (Feb. 4, 2020), https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [https://perma.cc/HN8B-DKVA]. A deepfake is a digitally altered "image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said." Mark Verstraete, *Inseparable Uses*, 99 N.C. L. Rev. 456, 459 n.144 (2021); *see also* J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 6.85 (April 2021); Matthew Bodi, *The First Amendment Implications of Regulating Political Deepfakes*, 47 Rutgers Comput. & Tech. L.J. 143, 144 (2021); Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1759 (2019) (noting that emergence of generative technology "will enable the production of altered . . . images, videos, and audios that are more realistic and more difficult to debunk than they have been in the past"); Richard L. Hasen, *Deep Fakes, Bots, and Siloed Justices: American Election Law in a "Post-Truth" World*, 64 ST. LOUIS U. L.J. 535, 542 (2002) (noting that deep fakes "audio and video clips can be manipulated using machine learning and artificial intelligence and can make . . . anyone else appear to say or do anything that the manipulator wants").

[49] Roth & Achuthan, *supra* note 48.

[50] *Synthetic and Manipulated Media Policy*, TWITTER HELP CTR., https://help.twitter.com/en/rules-and-policies/manipulated-media (last visited July 19, 2020) [https://perma.cc/MFF5-XMXM].

generally, Twitter labels content that is deceptively altered or fabricated and removes content if it impacts public safety or is likely to cause serious harm.[51] Twitter has already shown, on five separate occasions, that it will place warnings on posts from President Trump that violate its policies, including its manipulated media policy.[52]

In the first case of Twitter applying this new policy, Twitter labeled as "manipulated media" an edited video featuring then presidential candidate Joe Biden in which Biden appeared to be endorsing President Trump for re-election in 2020, which was tweeted by White House social media director Dan Scavino and retweeted by the President.[53] The video had been edited so as to mislead viewers into believing that Biden was actually endorsing Trump.[54]

[51] *See id.* Notably, media that meet all three of the criteria defined above — i.e., that are synthetic or manipulated, shared in a deceptive manner, and is likely to cause harm — may not be shared on Twitter and are subject to removal. Accounts engaging in repeated or severe violations of this policy may be permanently suspended. *Id.*

[52] Twitter's first warning labels on Tweets from the President involved unsubstantiated claims about mail-in ballots being fraudulent, glorifying violence/use of force, and a manipulated video (discussed further below). *See* Elizabeth Dwoskin, *Twitter's Decision to Label Trump's Tweets Was Two Years in the Making*, WASH. POST (May 29, 2020, 4:55 PM PDT), https://www.washingtonpost.com/technology/2020/05/29/inside-twitter-trump-label/ [https://perma.cc/N2ST-DE2Z]. Prior to his suspension, Twitter affixed a warning label to a second Tweet from the President promoting use of force against protestors, citing its policy regarding "the presence of a threat of harm against an identifiable group." Rachel Lerman, *Twitter Slaps Another Warning Label on Trump Tweet About Force*, WASH. POST (June 23, 2020, 3:34 PM PDT), https://www.washingtonpost.com/technology/2020/06/23/twitter-slaps-another-warning-label-trump-tweet-about-force/ [https://perma.cc/JS2Y-JKNP]. Facebook left the post up without a warning. *Id.*

[53] *See* Ivan Mehta, *Trump's Retweet with Doctored Biden Video Earns Twitter's First 'Manipulated Media' Label*, NEXT WEB (Mar. 9, 2020), https://thenextweb.com/twitter/2020/03/09/trumps-tweet-with-doctored-biden-video-earns-twitters-first-manipulated-media-label/ [https://perma.cc/K4BQ-UGN2].

[54] *Id.*

[Image taken from: Dan Scavino (@DanScavino), TWITTER (Mar. 7, 2020, 8:18 PM), https://twitter.com/DanScavino/status/1236461268594294785 [https://perma.cc/BQ95-RS4V].]

To better position itself to handle various types of misinformation relating to the process of voting and in anticipation of misinformation in connection with the announcement of election results, in May 2020 Twitter enacted and in September 2020 it expanded its Civic Integrity Policy.

Consistent with its general practice of removing only the most harmful speech while engaging in counterspeech with respect to less harmful speech, in accordance with its Civic Integrity Policy, Twitter adopted a policy of removing/censoring harmful tweets that may lead to interferences in, manipulation of, or intimidation regarding in the elections — such as tweets that encourage or threaten violence or call for people to interfere with election results or with the smooth operation of polling places — while responding with counterspeech, primarily in the form of labeling and reference to authoritative information by trusted sources, to tweets that it deemed less harmful

that implicated election integrity. Pursuant to this policy, Twitter committed to remove/censor attempts to manipulate or disrupt civic processes, "including through the distribution of false or misleading information about the procedures or circumstances around participation in a civic process,"[55] while responding with counterspeech to what it deems to be less harmful interferences in the election processes. Referencing this policy, Twitter explains: "In instances where misleading information does not seek to directly manipulate or disrupt civic processes, but leads to confusion . . . we may label the Tweets to give additional context" and may "reduce the visibility of Tweets containing false or misleading information about civic processes in order to provide additional context."

In addition, with respect to the announcement of post-election results, Twitter adopted a policy of generally engaging in counterspeech: Twitter's policy provides that

> [p]eople on Twitter, including candidates for office, may not claim an election win before it is authoritatively called. To determine the results of an election in the US, we require either an announcement from state election officials, or a public projection from at least two authoritative, national news outlets

[55] *Civic Integrity Policy*, TWITTER (Jan. 2021), https://help.twitter.com/en/rules-and-policies/election-integrity-policy [https://perma.cc/3ZVG-FDWL]. Twitter includes the following as examples of misleading information about participation in elections or other civic process:

- "misleading information about procedures to participate in a civic process (for example, that you can vote by Tweet, text message, email, or phone call . . .);
- misleading information about requirements for participation, including identification or citizenship requirements;
- misleading claims that cause confusion about the established laws, regulations, procedures, and methods of a civic process, or about the actions of officials or entities executing those civic processes; and
- misleading statements or information about the official, announced date or time of a civic process. . . .
- misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted;
- misleading claims about police or law enforcement activity related to voting in an election, polling places, or collecting census information;
- misleading claims about long lines, equipment problems, or other disruptions at voting locations during election periods;
- misleading claims about process procedures or techniques which could dissuade people from participating; and
- threats regarding voting locations or other key places or events . . . ."

Id.

> that make independent election calls. Tweets which include premature claims will be labeled and direct people to our official US election page.[56]

Twitter reserved for itself the discretion to determine whether to remove/censor or engage in counterspeech regarding misleading information about election outcomes, including the discretion to take some or all of the following measures: (1) apply a label and/or warning message to the content; (2) show a warning to people before they share or like the content; (3) reduce the visibility of the content on Twitter and/or preventing it from being recommended; and/or (4) provide a link to additional explanations or clarifications, such as in a "Twitter Moment" (a longer post than a tweet) produced by Twitter or a link to a Twitter policy (especially where the content is gaining significant attention or has caused substantial public confusion); and/or (5) restrict users' ability to reply, retweet, or like tweets.

In accordance with its Civic Integrity Policy, Twitter reserved for itself the right to take measures — either removal or counterspeech, depending on severity and source — in response to misleading information about election outcomes that was intended to undermine public confidence in the elections, including "disputed claims that could undermine faith in the process itself, such as unverified information about election rigging, ballot tampering, vote tallying, or certification of election results" and "misleading claims about the results or outcome of a civic process which calls for or could lead to interference with the implementation of the results of the process," for example, "claiming victory before election results have been certified, inciting unlawful conduct to prevent the procedural or practical implementation of election results."[57]

Twitter first exercised its discretion under the above policies to label misleading information by President Trump about mail-in voting. On October 26, 2020, President Trump issued a tweet claiming that there were "big problems and discrepancies with Mail In Ballots all over the USA." Twitter responded by engaging in the following types of counterspeech: (1) labeling the tweet as containing disputed content or as being potentially misleading, (2) creating a fact-check link below it providing accurate information about mail-in voting entitled "Voting by

---

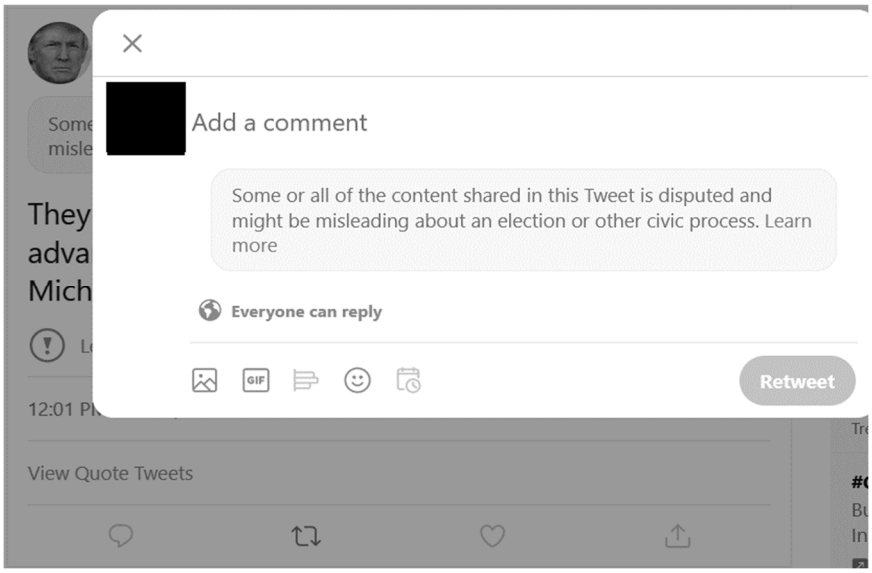
[56] Vijaya Gadde & Kayvon Beykpour, *Additional Steps We're Taking Ahead of the 2020 US Election*, TWITTER BLOG (Oct. 9, 2020), https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html [https://perma.cc/G3HT-MEAA] [hereinafter *Additional Steps*].

[57] *Civic Integrity Policy*, *supra* note 55.

mail is legal and safe, experts and data confirm," and (3) restricting Twitter users from liking, retweeting, or replying to the President's tweet. See below:

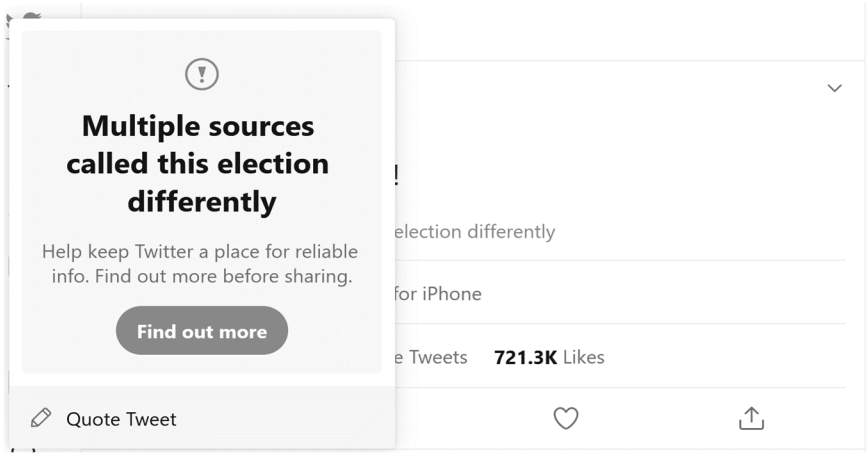
These counterspeech efforts were largely the same post-election. As of October 20, 2020, Twitter began setting its default to Quote Tweet rather than to Retweet, to encourage users to add their own counterspeech to the content they were interacting with before sharing it.[58]

[58] *See* Gadde & Beykpour, *Additional Steps*, *supra* note 56.



As opposed to Quote Tweeting, Retweeting was a single-click process for sharing content to a user's entire Twitter network.

Twitter provided the following prominent label when a user attempts to retweet (share) content with warnings like the above, and characterizes its approach as one of providing additional context:[59]



Twitter's actions included the labeling of eleven of twenty-two posts that President Trump made between November 3 and November 6.[60]

Anticipating a disputed election, on November 2, Twitter announced additional election integrity policies that expanded the criteria for Twitter to add counterspeech in the form of a label, as set forth below.[61]

Who is eligible for a label?

- All accounts with US 2020 candidate labels (including US 2020 Presidential candidate and campaign accounts)
- US-based accounts with more than 100,000 followers
- Tweets that have significant engagement (25k likes or 25k Quote Tweets and/or Retweets).

[59] *Cf.* Twitter Safety, *Expanding Our Policies to Further Protect the Civic Conversation*, TWITTER (Sept. 10, 2020), https://blog.twitter.com/en_us/topics/company/2020/civic-integrity-policy-update.html [https://perma.cc/X8R6-ULSF] (stating "non-specific, disputed information that could cause confusion about an election" should be accompanied by more context and Twitter will start to label any such tweets that "undermine public confidence in an election or other civic process").

[60] Rachel Sandler, *Half of Trump's Twitter and Facebook Posts Since Election Day Flagged*, FORBES (Nov. 6, 2020, 2:13 PM EST), https://www.forbes.com/sites/rachelsandler/2020/11/04/over-half-of-trumps-twitter-and-facebook-posts-since-election-day-flagged/ [https://perma.cc/RDX7-VT5J].

[61] *See* Gadde & Beykpour, *Additional Steps*, *supra* note 56.

Who do we consider official sources for election results?

- State election officials (as determined by the National Association of Secretaries of State and the National Association of State Election Directors)

National news outlets that have dedicated, independent election decision desks:

-ABC News
-Associated Press
-CBS News
-CNN
-Decision Desk HQ
-Fox News
-National Election Pool
-NBC News
-Reuters

The week after the election had been called, Twitter said it labeled 300,000 tweets related to the presidential election as disputed. Twitter also added a warning message and limited engagement features on 456 of those tweets.[62]

In the wake of the November 2020 election and of former president Trump's repeated false posts about the results of the election and his increasingly dangerous posts inciting the insurrection at the Capitol on January 6, 2021, Twitter adopted more aggressive counterspeech and removal policies. In the post-election period, Twitter attempted to slow down the rapidly proliferating misinformation about the result of the election, election fraud and related conspiracies, and "stop the steal" movement by carrying out previously announced election-integrity polices as well as by adding additional counterspeech and "friction" to the process of engaging with posts — and ultimately by removing the worst violations and violators from its platform. In the immediate post-election period, Twitter initially attempted to contain Trump's false election claims and the burgeoning "stop the steal" movement by flagging misinformation and restricting engagement; adding friction to

[62] Kate Conger, *Twitter Says It Labeled 0.2% Of All Election-Related Tweets As Disputed*, N.Y. TIMES (Nov. 12, 2020), https://www.nytimes.com/2020/11/12/technology/twitter-says-it-labeled-0-2-of-all-election-related-tweets-as-disputed.html [https://perma.cc/MRH6-D44Q]; Vijaya Gadde & Kayvon Beykpour, *An Update on Our Work Around the 2020 US Elections*, TWITTER BLOG (Nov. 12, 2020), https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html [https://perma.cc/A8DF-N4MH].

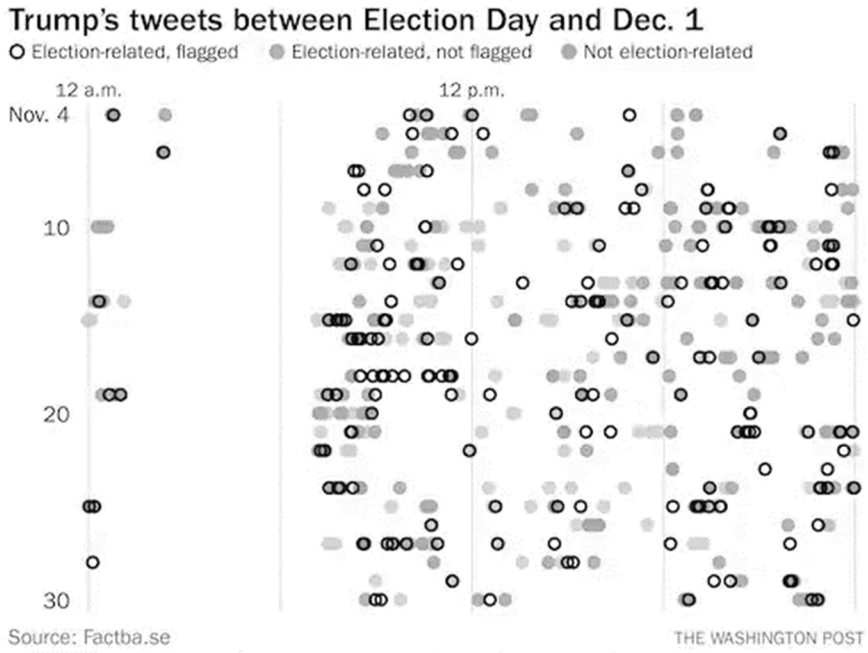
sharing posts to try to slow the spread of misinformation; flagging Trump's false claims and briefly limiting engagement; and ultimately by removing the worst offenders. Then, in the period immediately following the insurrection, Twitter took a number of additional, increasingly drastic measures, including the suspension of Trump for twelve hours, then eventually banning him permanently a day later; suspending related accounts that continued to spread election conspiracies, and purging QAnon affiliated accounts. I explore each of these measures in greater detail below.

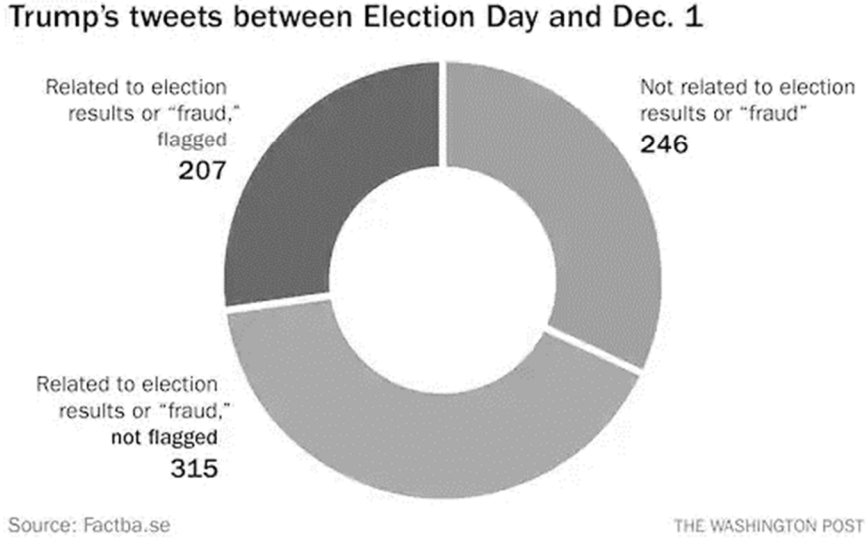
First, in the immediate wake of Election Day, Twitter placed a warning label on Trump's tweets embodying false claims of having won on Election Night,[63] including by imposing labels that required users to click in order to see the tweet, only permitting quote tweets, and disabling visible counts of retweets and likes.[64] By November 27, 2020, Twitter had flagged over 200 of Trump's posts as disputed or misleading — which amounted to about thirty percent of all of his posts since Election Day.[65]

[63] @TwitterSafety, TWITTER (Nov. 4, 2020, 1:04 AM), https://twitter.com/TwitterSafety/status/1323868590047744000 [https://perma.cc/DC7Y-BSKC].

[64] Geoffrey A. Fowler, *Twitter and Facebook Warning Labels Aren't Enough to Save Democracy*, WASH. POST (Nov. 9, 2020), https://www.washingtonpost.com/technology/2020/11/09/facebook-twitter-election-misinformation-labels/ [https://perma.cc/6KPW-N598]. Commenters note how Twitter is more willing to experiment with product changes to slow the spread of information despite disruptions to its service than Zuckerberg has expressed Facebook would be. Zuckerberg is quoted as saying, "[o]nce we're past these events, and we've resolved them peacefully, I wouldn't expect that we continue to adopt a lot more policies that are restricting of a lot more content." *Id.*

[65] Todd Spangler, *Twitter Has Flagged 200 of Trump's Posts as 'Disputed' or Misleading Since Election Day. Does It Make a Difference?*, VARIETY (Nov. 27, 2020, 8:54 AM PT), https://variety.com/2020/digital/news/twitter-trump-200-disputed-misleading-claims-election-1234841137/ [https://perma.cc/TEB8-QSCP].





See charts from Washington Post.[66]

[66] Philip Bump, *Twitter Keeps Flagging Trump for Disinformation Because Trump Keeps Tweeting Disinformation*, WASH. POST (Dec. 2, 2020, 8:17 AM PST), https://www.washingtonpost.com/politics/2020/12/02/twitter-keeps-flagging-trump-

In December, amidst the deluge of misinformation coming from @realdonaldtrump, Twitter added restrictions to how users could engage with three of Trump's flagged tweets. In particular, Twitter prevented users from liking, retweeting, replying, and copying the URL. Additionally, counts were disabled and while quote tweets were permitted after clicking through a warming pop-up, they were undiscoverable.[67]

Twitter also temporarily adopted an initiative (which it terminated on December 16, 2020) that was intended to encourage quote tweets by adding some "friction" through a prompt that appeared when a user went retweet a tweet.[68] Twitter intended for this approach to slow down the spread of misinformation by adding an extra step.[69]

Then, on January 6, 2021, in the immediate wake of the insurrection after Trump's tweets condoning and encouraging the violence the Capitol, Twitter responded to these developments initially by imposing a twelve-hour suspension on President Trump and by warning him of a permanent ban in the case of further violations.[70] On January 8, Twitter announced the permanent suspension of @realDonaldTrump after Trump posted tweets sympathizing with the insurrectionists and announcing that he would not attend the Inauguration of President Biden.[71] Twitter premised this ban on its determination that Trump's tweets could reasonably be interpreted as encouraging further violence in relation to the Inauguration. In addition, Twitter banned several Trump-affiliated accounts that Twitter determined were contributing to

disinformation-because-trump-keeps-tweeting-disinformation/ [https://perma.cc/BS6A-HP4E].

67 Kim Lyons, *Twitter Briefly Restricts Trump's Disputed Election Tweets*, VERGE (Dec. 12, 2020, 10:43 AM EST), https://www.theverge.com/2020/12/12/22171126/trump-twitter-disputed-tweets-election-retweets [https://perma.cc/GNJ8-GRHM]. A Twitter spokesperson said this was inadvertent and the platform reverted back to simply labeling misinformation. *Id.*

68 *See id.*

69 *See id.*

70 Bobby Allyn, *Twitter Locks Trump's Account, Warns of 'Permanent Suspension' if Violations Continue*, NPR (Jan. 6, 2021, 7:53 PM ET), https://www.npr.org/sections/congress-electoral-college-tally-live-updates/2021/01/06/954190994/twitter-locks-trumps-account-warns-of-permanent-suspension-if-violations-continue [https://perma.cc/G8U8-CUL2].

71 Twitter Inc., *Permanent Suspension of @realDonaldTrump*, TWITTER (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html [https://perma.cc/72CF-WASV].

and supporting the insurrection (by sharing conspiracy theories, etc.) and also banned Trump from using the @POTUS account.[72]

In the days following the January 6 insurrection, Twitter additionally reported that it purged over 70,000 accounts affiliated with QAnon[73] (which incidentally resulted in nearly every major GOP elected official losing a significant number of followers).[74]

---

[72] *See, e.g.*, Jack Brewster, *Lin Wood — Lawyer Closely Tied to Trump — Permanently Banned from Twitter After Claiming Capitol Siege Was 'Staged*,' FORBES (Jan. 7, 2021, 1:42 PM EST), https://www.forbes.com/sites/jackbrewster/2021/01/07/lin-wood-lawyer-closely-tied-to-trump-permanently-banned-from-twitter-after-claiming-capitol-siege-was-staged/ [https://perma.cc/A6AL-ZBFL] (permanent suspension of Lin Wood, @LlinWood, on Jan. 7, 2021 for violating rules against inciting violence and @FightBackLaw, an account Mr. Wood used to attempt to evade the ban); Brakkton Booker, *My Pillow CEO Mike Lindell Permanently Suspended from Twitter*, NPR (Jan. 26, 2021, 10:17 AM ET), https://www.npr.org/2021/01/26/960679189/my-pillow-ceo-mike-lindell-permanently-suspended-from-twitter [https://perma.cc/WFZ2-4DQ2] (permanent suspension of Mike Lindell, @realMikeLindell, on Jan. 25, 2021 for repeat violations of Twitter's Civic Integrity Policy); Bill Chappell, *Twitter Suspends Rep. Marjorie Taylor Greene's Account*, NPR (Jan. 17, 2021, 5:22 PM ET), https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/17/957891462/twitter-suspends-rep-marjorie-taylor-greene-s-account-temporarily [https://perma.cc/Q597-EHYR] (temporary suspension of Marjorie Taylor Green, @mtgreenee, on Jan. 17, 2021); Ben Collins & Brandy Zadrozny, *Twitter Bans Michael Flynn, Sidney Power in Qanon Account Purge*, NBC NEWS (Jan. 8, 2021, 1:28 PM PST), https://www.nbcnews.com/tech/tech-news/twitter-bans-michael-flynn-sidney-powell-qanon-account-purge-n1253550 [https://perma.cc/3F5M-UKGZ] (permanent suspension of Michael Flynn, @GenFlynn, on Jan. 8, 2021 for sharing Qanon content); *id.* (permanent suspension of Sidney Powell, @SidneyPowell1, on Jan. 8, 2021 for sharing QAnon content); Lindsey Ellefson, *Fox News' Dan Bongino Won't Return to Twitter After Suspension: 'F– You*,' WRAP (Jan. 7, 2021, 2:12 PM), https://www.thewrap.com/dan-bongino-quits-twitter-after-suspension/ [https://perma.cc/L9U5-V39E] (temporary suspension of Dan Bongino, @dbongino, on Jan. 7, 2021 for violating Twitter's Civic Integrity Policy); Sean Hollister, *Twitter Is Deleting Trump's Attempts to Circumvent Ban*, VERGE (Jan. 8, 2021, 9:50 PM EST), https://www.theverge.com/2021/1/8/22221683/trump-tried-to-evade-his-ban-with-potus-but-those-tweets-were-instantly-deleted [https://perma.cc/DLF3-KP7Q] (permanent suspension of Team Trump, @TeamTrump, on Jan. 8, 2021 for being used by Trump to attempt to evade his ban); Zen Soo, *Twitter Permanently Bans My Pillow CEO*, AP NEWS (Jan. 26, 2021), https://apnews.com/article/joe-biden-donald-trump-media-elections-presidential-elections-ac34de7cb5844d96589a10ea6e653d50 [https://perma.cc/7ENE-DA4F] (permanent suspension of My Pillow, @mypillowusa, on Feb. 1, 2021 for violating Twitter's policy on ban evasion because Mike Lindell was using it to post).

[73] Tony Romm & Elizabeth Dwoskin, *Twitter Purged More than 70,000 Accounts Affiliated with Qanon Following Capitol Riot*, WASH. POST (Jan. 11, 2021, 6:57 PM PST), https://www.washingtonpost.com/technology/2021/01/11/trump-twitter-ban/ [https://perma.cc/B9RV-NQKJ].

[74] Some notable Trump affiliates who lost large amounts of followers include House Minority Leader Kevin McCarthy, U.S. Secretary of State Mike Pompeo, Former Senate Majority Leader Mitch McConnell, Rep. Clay Higgins, who lost fifteen percent of his entire Twitter following, Rep. Devin Nunez, who lost fifteen percent of his entire

In the post-insurrection period, Twitter updated its policies to target more deliberately those spreading election conspiracy theories. Twitter's January 12, 2021 safety policy updates included the following measures:

- Continued/heightened monitoring and reducing the visibility of those who have posted or engaged with QAnon or other coordinated harmful activity.
- Limited engagement on tweets that have been labeled for violating Twitter's civic integrity policy.
- Prevention of certain content from trending, including tweets with terms that violate Twitter's rules regarding Coordinated Harmful Activity, Civic Integrity, Hateful Conduct, Glorification of Violence, Violent Threats, and/or Sensitive Media. [75]

Finally, in connection with the Inauguration on January 20, 2021, Twitter adopted certain counterspeech measures by creating an official inauguration hub populated by coverage from reliable information sources.[76]

In short, in the leadup to and in the aftermath of the 2020 presidential election, Twitter adopted an increasingly aggressive counterspeech policy to combat political and election-related misinformation on its

Twitter following, Sen. Kelly Loeffler, who lost ten percent of her entire Twitter following, Rep. Jim Jordan, Sen. Rand Paul, Rep. Dan Crenshaw, Sen. Ted Cruz, Rep. Matt Gaetz, Rep. Doug Collins, Sen. Chuck Grassley, Sarah Huckabee Sanders, Dave Rubin, Bari Weiss, and more. Angela Wang, *GOP Politicians Lost Tens of Thousands of Followers After Twitter Purged Qanon Accounts, Here's Who Lost the Most.*, BUS. INSIDER (Jan. 14, 2021, 6:00 PM), https://www.businessinsider.com/gop-officials-lost-most-twitter-followers-qanon-purge-2021-1 [https://perma.cc/JSU2-WNZ2].

[75] Twitter Safety, *An Update Following the Riots in Washington, DC*, TWITTER BLOG (Jan. 12, 2021), https://blog.twitter.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--.html [https://perma.cc/Y5ZR-2AF7].

[76] @TwitterGov, *What to Expect on Twitter on US Inauguration Day 2021*, TWITTER BLOG (Jan. 14, 2021), https://blog.twitter.com/en_us/topics/company/2021/inauguration-2021.html [https://perma.cc/NJ5E-W973]. In addition, on Inauguration Day, Twitter carried out its usual transferring of official White House Twitter accounts and archiving of former accounts. Posts from @realdonaldtrump have disappeared from Twitter and are now only accessible through third-party archives. *See* David Yanofsky, *Where to Read Donald Trump's Tweets Now that Twitter Has Closed His Account*, QUARTZ (Jan. 8, 2021), https://qz.com/1955036/where-to-find-trumps-tweets-now-that-hes-banned-from-twitter/ [https://perma.cc/UBD4-669K]; *see also* Politwoops, *Deleted Tweets from Donald J. Trump, R-Fla.*, PROPUBLICA, https://projects.propublica.org/politwoops/user/realDonaldTrump (last visited Feb. 25, 2021) [https://perma.cc/TCB5-V82Y].

platform. Twitter then pivoted to a removal/censorship approach only when it determined that the speech of former President Trump (and that of associated and like-minded speakers) was likely to cause imminent real-world violence.

Twitter continues to experiment with novel approaches to combatting misleading and harmful speech on its platform. In late January 2021, Twitter introduced "Birdwatch," which it describes as "a community-based approach to misinformation" and which provides a new vehicle for facilitating counterspeech and combating harmful speech on its platform.[77] Birdwatch allows Twitter to crowdsource the problem of misinformation on its platform by enabling ordinary users to engage in counterspeech in the form of writing "notes" in response to content a user believes is misleading.[78] In response, other Twitter users can rate whether such a "note" is helpful, which in turn factors into Twitter's determination of the note's level of credibility and visibility.[79] Such notes will eventually "travel with" the tweets they are commenting on, so that other Twitter users can see tweets and corresponding notes side-by-side.[80] Twitter intends for this crowdsourcing effort to help shore up and rebuild trust in its platform — and its platform's fact-checking initiatives — by allowing users themselves to be a part of the fact-checking process.[81] As of this writing, the pilot version of the program in the United States has about 2,000 participants[82] and is visible only on a special Twitter Birdwatch site at

[77] Keith Coleman, *Introducing Birdwatch, a Community-Based Approach to Misinformation*, TWITTER BLOG (Jan. 25, 2021), https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html [https://perma.cc/82UV-ZX9X].

[78] *Id.*

[79] *Id.*

[80] *Id.*; *see also FAQs*, BIRDWATCH GUIDE, https://twitter.github.io/birdwatch/about/faq/ [https://perma.cc/Q6P2-8K27] (last visited Apr. 14, 2021) ("Eventually, we aim to make notes visible directly on Tweets for the global Twitter audience when there is consensus from a broad and diverse set of contributors.")

[81] Kurt Wagner, *Inside Twitter's Plan to Fact-Check Tweets*, BLOOMBERG (Mar. 4, 2021, 3:45 AM PST), https://www.bloomberg.com/news/newsletters/2021-03-04/birdwatch-inside-twitter-s-plan-to-fact-check-tweets [https://perma.cc/MEV9-3HHW] ("Trust in the process and the way this is done is the biggest motivator behind Birdwatch . . . . We very consistently heard across the political spectrum people saying that they felt like they would value a community-driven approach, in many cases more than what Twitter does today.").

[82] Elizabeth Culliford, *Twitter's Birdwatch Crowd Experiment Courts Familiar Challenges*, REUTERS (Mar. 19, 2021, 4:05 PM), https://www.reuters.com/article/us-twitter-moderation-birdwatch-focus/twitters-birdwatch-crowd-experiment-courts-familiar-challenges-idUSKBN2BB13A [https://perma.cc/M7QY-LLYA].

birdwatch.twitter.com, not on the main Twitter site.[83] Twitter faces a number of challenges in implementing this novel and ambitious process for facilitating counterspeech on its platform, including ensuring that the note-writers represent a broad and diverse cross-section of perspectives from the Twitter community and that the process is not taken over by coordinated manipulation attempts or other types of abuse or harassment.[84]

In sum, consistent with First Amendment values, Twitter generally opts for responses and interventions in the form of counterspeech (such as labeling, warnings interposed in users' processes of sharing content, references to authoritative sources, and the crowdsourcing of counterspeech via the Birdwatch program) instead of censorship — and such counterspeech proved to be moderately effective in reducing the spread of such misinformation. Notable exceptions include removal of tweets calling for violence or interference at the polls, and manipulated media impacting public safety or likely to cause serious harm, and ultimately the permanent ban on former President Trump, through which Twitter has prohibited Trump from ever using its platform again — regardless of whether Trump is ever re-elected — as a consequence of Trump's speech inciting violence in the context of the insurrection at the Capitol.

B. Facebook

Facebook has also adopted a host of policies to address potentially harmful political and election-related misinformation. First, Facebook adopted policies regarding misinformation about the voting process and the post-election announcement of election results. Second, Facebook adopted policies regarding misinformation that is applicable to political and campaign related speech generally.

In preparation for the 2020 presidential election, Facebook adopted policies to address voter suppression and intimidation, as well as policies regarding election results announcements.[85] Pursuant to these policies, Facebook removed content that was directed to suppressing votes or intimidating voters, including posts that contained any of the following false assertions of fact (regardless of their source):

---

[83] BIRDWATCH, https://twitter.com/i/birdwatch (last visited Apr. 14, 2021) [https://perma.cc/CUZ3-8MHT].

[84] BIRDWATCH GUIDE, *supra* note 80.

[85] Guy Rosen, *Preparing for Election Day*, FACEBOOK (Oct. 7, 2020), https://about.fb.com/news/2020/10/preparing-for-election-day/ [https://perma.cc/87RQ-W5VS].

- Misrepresentation of the dates, locations, times and methods for voting or voter registration (e.g., "Vote by text!");
- 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 (e.g., "If you voted in the primary, your vote in the general election won't count."); and
- Threats of violence relating to voting, voter registration or the outcome of an election.[86]

In addition, Facebook adopted policies under which it bans advertising that suggests that voting is useless or meaningless or advises people not to vote. Regarding election result announcements, Facebook adopted a policy of directing users to its Voting Information Center for real-time vote-counting results and applying warnings to posts claiming election victory prematurely.[87]

Applying its voter suppression and intimidation policies, Facebook responded with counterspeech/labeling to eleven of twenty-two posts President Trump made between November 3 and November 6 regarding the outcome of the election. In addition, on November 5, Facebook shut down a group with over 300,000 users called "Stop the Steal" that was organizing protests claiming Joe Biden was trying to steal the election. Facebook labeled these posts and referred users to accurate and truthful information regarding the outcome of the election.

Facebook's labeling and other counterspeech measures were apparently somewhat less effective than those of Twitter. Third-party estimates of the efficacy of Facebook's approach suggest that it led to only an eight percent decrease in the sharing of this false post.[88] By comparison, Twitter's warning of misinformation prior to permitting sharing reduced shares by nearly thirty percent, as noted above, and Facebook's own practice of fact-checking, as opposed to merely directing users to authoritative content, reduced the spread of content

[86] *Id.*

[87] *Id.*

[88] Brianna Provenzano, *Facebook Knows that Labeling Trump's Election Lies Hasn't Stopped His Posts from Going Viral*, GIZMODO (Nov. 16, 2020, 8:24 PM), https://gizmodo.com/facebook-knows-that-labeling-trumps-election-lies-I-1845693925 [https://perma.cc/JAR8-22S6].

by eighty percent once labeled as false. As a matter of company policy, Facebook does not fact-check politicians.[89]

Facebook has taken a number of steps to combat misinformation in general on its platform.[90] Consistent with its general approach of favoring counterspeech/labeling over censorship/removal, Facebook's efforts to combat misinformation have trended toward labeling and fact-checking, rather than removal. The company has adopted extensive measures to attempt to combat publicly-available misinformation on its platform, including by partnering with independent third-party fact-checkers to evaluate posts and providing counterspeech in the form of "Related Articles"/"Additional Reporting on This" on topics similar to false or misleading posts.[91] These extensive measures to combat misinformation and false content on Facebook are generally applicable to political content and political ads, but are not applicable to posts that are considered "direct speech by a politician."[92] Thus, under Facebook's currently applicable fact-checking policies, political speech and the content of political ads are subject to fact-checking — except if such content constitutes "direct speech by a politician."[93] This exception for politicians' content has come under substantial scrutiny in recent months, especially given the highly controversial posts of President Trump.[94] Before examining this controversial exception to Facebook's general labeling/fact-checking/counterspeech policy for public posts on

[89] Craig Silverman & Ryan Mac, *Facebook Knows that Adding Labels to Trump's False Claims Does Little to Stop Their Spread*, BUZZFEED (Nov. 16, 2020, 8:07 PM ET), https://www.buzzfeednews.com/article/craigsilverman/facebook-labels-trump-lies-do-not-stop-spread [https://perma.cc/GP2T-ST85].

[90] *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 [https://perma.cc/9D5P-7CBA] [hereinafter *Facebook's Strategy*].

[91] *See, e.g.*, Hunt Allcott, Matthew Gentzkow & Chuan Yu, *Trends in the Diffusion of Misinformation on Social Media Online Appendix* 4 (Stanford Inst. for Econ. Policy Research, Working Paper No. 18-029, 2018), http://web.stanford.edu/~gentzkow/research/fake-news-trends-appx.pdf [https://perma.cc/Y2X9-D276] (listing in Table 1 all of Facebook's efforts to combat false news).

[92] *Program Policies*, FACEBOOK BUS. HELP CTR., https://www.facebook.com/business/help/315131736305613 (last visited Sept. 4, 2020) [https://perma.cc/8JWX-WTFA].

[93] *Id.*

[94] *See* Michael M. Grynbaum & Tiffany Hsu, *CNN Rejects 2 Trump Campaign Ads, Citing Inaccuracies*, N.Y. TIMES (Oct. 3, 2019), https://www.nytimes.com/2019/10/03/business/media/cnn-trump-campaign-ad.html [https://perma.cc/2TSS-XXU9]; Cecilia Kang, *Facebook's Hands-Off Approach to Political Speech Gets Impeachment Test*, N.Y. TIMES (Oct. 8, 2019), https://www.nytimes.com/2019/10/08/technology/facebook-trump-biden-ad.html [https://perma.cc/AM92-AC2D].

its platform, I first examine the company's generally-applicable policy itself.

Over the past four years, Facebook has expanded the partnership it began in December 2016 with fact-checkers to evaluate publicly-available content posted on its platform.[95] Through its fact-checking initiatives, Facebook works with select independent third-party fact-checkers, which are certified through the non-partisan International Fact-Checking Network.[96] In the United States, the certified fact-checking organizations with whom Facebook works are the Associated Press, factcheck.org, Lead Stories, Check Your Fact, Science Feedback, and PolitiFact.[97] Facebook has expanded its general fact-checking initiative to include the fact-checking of all public, newsworthy Facebook posts, including links, articles, photos, and videos.[98] The fact-checking process on Facebook applies to political advertisements unless those advertisements (or other posts) constitute "direct speech made by an elected official."[99] The fact-checking process can be initiated by Facebook users who flag a post as being potentially false. Subject to the exception for direct speech by politicians, any public, newsworthy post (including text, photos, and videos) can be flagged for fact-checking, either by a user, by an outside journalist, or, as is most commonly the case, by Facebook's machine learning algorithms. For a user to flag a post as potentially false, a user clicks "•••" next to the post he or she

[95] *See* Lyons, *Facebook's Strategy*, *supra* note 90.

[96] *Id.*; *see also Verified Signatories of the IFCN Code of Principles*, POYNTER, https://ifcncodeofprinciples.poynter.org/signatories (last visited Sept. 12, 2020) [https://perma.cc/JL6K-DGJE].

[97] *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 [https://perma.cc/R5YC-ZVNT]; *see also Fact-Checking on Facebook*, FACEBOOK HELP CTR., https://www.facebook.com/help/publisher/182222309230722 (last visited July 19, 2020) [https://perma.cc/LQ5G-W38R] (providing an overview of Facebook's fact-checking program); *How Are Independent Fact-Checkers Selected on Facebook?*, FACEBOOK HELP CTR., https://www.facebook.com/help/1599660546745980 (last visited Sept. 29, 2018) [https://perma.cc/M5UK-Q2C7] (explaining how a third-party becomes a fact-checker for Facebook). Notably, 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 politically biased, but this publication is now defunct.

[98] *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 [https://perma.cc/9V6C-LHPH].

[99] "If a claim is made directly by a politician on their Page, in an ad or on their website, it is considered direct speech and ineligible for our third party fact checking program — even if the substance of that claim has been debunked elsewhere." *See Program Policies*, *supra* note 92.

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."[100]

Once a post is flagged by a user, journalist, or Facebook's machine learning as a potential false news story, it is submitted for evaluation to a third-party independent fact-checker.[101] For each piece of content up for review, a fact-checker has the option of providing one of six different ratings: false, altered, partly false, missing context, satire, or true.[102] If 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 activity, and likes), such that future views of each false post will be reduced by an average of eighty percent.[103] Second, Facebook may commission 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.[104] Such content is then displayed in conjunction with the false post on the same subject.[105] While Facebook formerly flagged

[100] *See How Do I Mark a Facebook Post as False News?*, FACEBOOK HELP CTR., https://www.facebook.com/help/572838089565953 (last visited Sept. 29, 2018) [https://perma.cc/VSQ3-SRGR]. Alternatively, a user can click "•••" next to a post, then click "Find Support or Report Post," and then select "False News." *Id.*

[101] *See* Tessa Lyons, *Hard Questions: How Is Facebook's Fact-Checking Program Working?*, FACEBOOK NEWSROOM (June 14, 2018), https://about.fb.com/news/2018/06/hard-questions-fact-checking/ [https://perma.cc/SD33-3YWT] [hereinafter *Facebook's Fact-Checking*] ("[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.").

[102] *Rating Options for Fact-Checkers*, FACEBOOK BUS. HELP CTR., https://www.facebook.com/business/help/341102040382165 (last visited Aug. 28, 2020) [https://perma.cc/4ZKY-BVWW].

[103] Lyons, *Facebook's Fact-Checking*, *supra* note 100; *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/ [https://perma.cc/Z9GK-GAJ2].

[104] *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 [https://perma.cc/DA48-UBPE] [hereinafter *Disputed Flags*].

[105] Geoffrey A. Fowler, *I Fell for Facebook Fake News. Here's Why Millions of You Did Too.*, WASH. POST (Oct. 18, 2018, 1:58 PM PDT), https://www.washingtonpost.com/technology/2018/10/18/i-fell-facebook-fake-news-heres-why-millions-you-did-too/ [https://perma.cc/YJ9T-DF64] (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, *Disputed Flags*, *supra* note 104; *see also* Sara Su, *New Test with Related Articles*, FACEBOOK NEWSROOM (Apr. 25, 2017),

false news sites with a "Disputed" flag, the company is experimenting with different approaches in response to research suggesting that such flags may actually entrench beliefs in the disputed posts.[106] Facebook now provides "Related Articles"/"Additional Reporting on This" in conjunction with false news stories, which apparently does not result in similar entrenchment.[107] 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,[108] as in the example below (setting forth Facebook and Instagram's flags).[109]

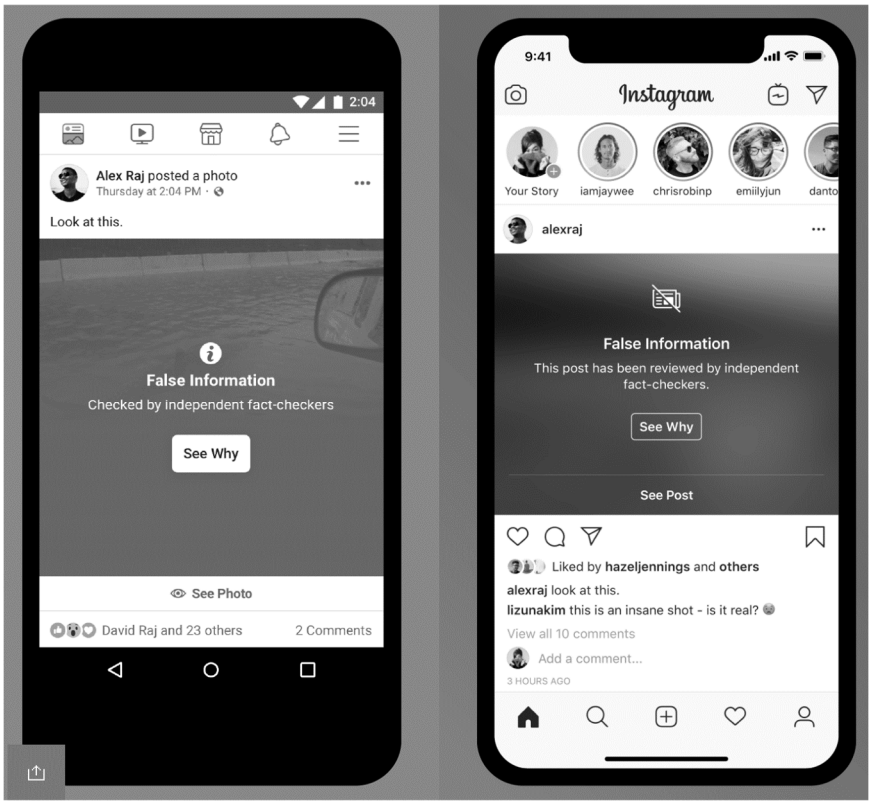
https://newsroom.fb.com/news/2017/04/news-feed-fyi-new-test-with-related-articles [https://perma.cc/7SR6-L3H5].

106 *See* Lyons, *Disputed Flags*, *supra* note 104.

107 *See id.* (explaining that "[a]cademic 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 . . . [but that] 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").

108 *See id.*

109 *E.g.*, Elle Hunt, *'Disputed by Multiple Fact-Checkers': Facebook Rolls Out New Alert to Combat Fake News*, GUARDIAN (Mar. 21, 2017, 8:37 PM EDT), https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news [https://perma.cc/L4KM-L925].



[Image taken from: Karissa Bell, *Instagram adds 'false information' labels to prevent fake news from going viral*, MASHABLE (Oct. 21, 2019), https://mashable.com/article/instagram-false-information-labels/ [https://perma.cc/H758-4QUV].]

In addition, as Facebook explains: "When fact-checkers write articles with more information about a story, you'll see a notice where you can click to see why."[110] Facebook also provides its users who are about to share posts that have been debunked by a fact-checker by alerting them to additional reporting.[111] Facebook also now posts more prominent fact-checking labels as interstitials warnings atop photos and videos on Facebook (and Instagram) that were fact-checked as false.

Facebook's general false news policy, composed of the fact-checking process and counterspeech mechanisms described above, is not

[110] *How Is Facebook Addressing False Information Through Independent Fact-Checkers?*, FACEBOOK HELP CTR., https://www.facebook.com/help/1952307158131536 (last visited July 21, 2020) [https://perma.cc/7BQ8-BZ2V].

[111] *Id.*

applicable to "direct speech" by politicians. Such direct speech by politicians is not run through Facebook's external fact-checking process nor subject to labeling or the commissioning of counterspeech in response.[112] Facebook proffers the following justification for this exception to its fact-checking policy:

> We rely on third-party fact-checkers to help reduce the spread of false news and other types of viral misinformation, like memes or manipulated photos and videos. We don't believe, however, that it's an appropriate role for us to referee political debates and prevent a politician's speech from reaching its audience and being subject to public debate and scrutiny . . . . This means that *we will not send organic content or ads from politicians to our third-party fact-checking partners for review*.[113]

Posts and ads that constitute "direct speech" from current "politicians" at any/every level and their appointees — i.e., the politician's own claim or statement — are not subjected to fact-checking — even if the substance of the claim has been debunked elsewhere.[114]

Facebook's decision not to submit direct speech from current politicians to fact-checking is apparently grounded in the belief that such political speech is already subject to sufficient scrutiny among the polity and the free press and should not be subject to further scrutiny by Facebook's fact-checkers.[115] Facebook further justifies its policies as follows: "In a democracy, people should decide what is credible, not tech companies . . . . That's why - like other internet platforms and broadcasters - we don't fact check ads from politicians."[116] As a result, political speech and political ads made by politicians themselves — posts and campaign ads by politicians — operate in a separate system on Facebook.

Facebook's decision not to screen for or remove false posts or ads by politicians came into sharp focus in October 2019, when President Donald Trump's reelection campaign began running an ad that was

---

[112] *See Program Policies*, *supra* note 92.

[113] Clegg, *Facebook, Elections and Political Speech*, *supra* note 42 (emphasis added). Facebook will, however, subject the posts of political action committees and political advocacy groups to its fact-checking process. Klepper, *supra* note 41.

[114] Kate Cox, *Political Ads Can Lie If They Want, Facebook Confirms*, ARS TECHNICA (Oct. 10, 2019, 9:55 AM), https://arstechnica.com/tech-policy/2019/10/political-ads-can-lie-if-they-want-facebook-confirms/ [https://perma.cc/P8N7-HYWT]; *Fact-Checking on Facebook*, *supra* note 97.

[115] *Program Policies*, *supra* note 92.

[116] Klepper, *supra* note 41 (quoting Facebook's company statement).

proven to be false about then presidential candidate Joe Biden on Facebook.[117] The Trump Campaign released a thirty-second video ad accusing Biden of promising Ukraine money in exchange for firing a prosecutor investigating a company with ties to Biden's son, Hunter Biden.[118] The video ad falsely claimed that Joe Biden offered Ukraine $1 billion in aid if Ukraine pushed out the official investigating a company tied to Hunter Biden.[119] The Biden campaign asked Facebook to take down the ad, but Facebook refused.[120] In justifying its refusal, Facebook's head of global elections policy Katie Harbath explained: "Our approach is grounded in Facebook's fundamental belief in free expression, respect for the democratic process, and the belief that, in mature democracies with a free press, political speech is already arguably the most scrutinized speech there is."[121] Accordingly, the false Trump Campaign ad on Biden remained available on Facebook.

Facebook has encountered strong opposition to its policy exempting politicians' (and especially President Trump's) posts from fact-checking and from other of the company's content policies as well, including those prohibiting threats of imminent violence. One particular flashpoint at issue involved violent speech in the form of Donald Trump's May 2020 post following the murder of George Floyd and the ensuing demonstrations.[122] Trump threatened to deploy the military in Minneapolis to "bring the City under control" and infamously stated "when the looting starts, the shooting starts."[123]

---

[117] *See* Amy Sherman, *Donald Trump Ad Misleads About Joe Biden, Ukraine, and the Prosecutor*, POLITIFACT (Oct. 11, 2019), https://www.politifact.com/factchecks/2019/oct/11/donald-trump/trump-ad-misleads-about-biden-ukraine-and-prosecut/ [https://perma.cc/UY87-LN9S].

[118] Grynbaum & Hsu, *supra* note 94.

[119] *Id.*

[120] Kang, *supra* note 94.

[121] *Id.*

[122] *See* Megan Rose Dickey & Taylor Hatmaker, *Facebook Employees Stage Virtual Walkout in Protest of Company's Stance on Trump Posts*, TECHCRUNCH (June 1, 2020, 10:01 AM PDT), https://techcrunch.com/2020/06/01/facebook-employees-stage-virtual-walkout-in-protest-of-companys-stance-on-trump-posts/ [https://perma.cc/64V4-482Y].

[123] *Id.*

[Image taken from: Donald J. Trump (@realDonaldTrump), TWITTER (May 29, 2020, 12:53 AM), https://twitter.com/realDonaldTrump/status/1266231100780744704 [https://perma.cc/9DCM-9F7G].]

President Trump made this post across several platforms. While Twitter appended a notice to the President's post explaining that the post violated the platform's rules against glorifying violence and requiring users to click through the notice to view the tweet (see below), Facebook took no action.[124]

[124] *See* Brian Stelter & Donie O'Sullivan, *Trump Tweets Threat that 'Looting' Will Lead to 'Shooting.' Twitter Put a Warning Label on It*, CNN (May 29, 2020, 10:40 AM ET), https://www.cnn.com/2020/05/29/tech/trump-twitter-minneapolis/index.html [https://perma.cc/W7JV-YE62] (including a screenshot of the notice Twitter posted in response to Trump's tweet).

[Image taken from: Twitter Comms (@TwitterComms), TWITTER (May 29, 2020, 3:17 AM), https://twitter.com/TwitterComms/status/1266267447838949378 [https://perma.cc/6MZD-BYHA].]

Facebook's CEO Mark Zuckerberg explained that he was personally appalled by the President's tweet, but felt that Facebook's institutional role was to "enable as much expression as possible unless it will cause imminent risk of specific harms or dangers spelled out in [Facebook's] clear policies."[125] Some of Facebook's employees, however, voiced dissatisfaction with the company's response.[126] Facebook ultimately retreated from its non-interventionist stance towards Donald Trump and his campaign, at least with respect to its hate speech content regulation, and subsequently removed a Trump Campaign page ad because of its use of a symbol of hate.[127] In addition, Facebook recently

[125] Mark Zuckerberg, FACEBOOK (May 29, 2020), https://www.facebook.com/zuck/posts/10111961824369871 [https://perma.cc/7JQX-KRGT].

[126] Rachel Siegel & Elizabeth Dwoskin, *Facebook Employees Blast Zuckerberg's Hands-Off Response to Trump Posts as Protests Grip Nation*, WASH. POST (June 1, 2020, 5:04 PM PDT), https://www.washingtonpost.com/business/2020/06/01/facebook-zuckerberg-donation-trump/ [https://perma.cc/Q7SE-Q9JG].

[127] Isaac Stanley-Becker, *Facebook Removes Trump Ads with Symbol Once Used by Nazis to Designate Political Prisoners*, WASH. POST (June 18, 2020, 12:48 PM PDT), https://www.washingtonpost.com/politics/2020/06/18/trump-campaign-runs-ads-with-marking-once-used-by-nazis-designate-political-prisoners/ [https://perma.cc/Y96D-M5MS]. Days later when a Trump-affiliated campaign page posted an advertisement denouncing "dangerous MOBS" accompanied by an image of a downward facing red

announced that it would "remove posts [from political leaders] that incite violence or attempt to suppress voting . . . [and] affix labels on posts that violate hate speech prohibitions."[128]

Facebook's decision to exempt speech by politicians from its fact-checking and other content regulation policies also drew sharp criticism recently. Civil rights and liberties leader Laura W. Murphy, along with a team from civil rights law firm Relman Colfax, conducted an extensive, independent two-year civil rights audit of Facebook's content regulation policies and their implementation.[129] The auditors' concerns were magnified by Facebook's response to President Trump's posts regarding recent civil rights protests and mail-in ballots in the context of the pandemic.[130] The auditors expressed strong criticisms of the company's policies and exemption of Trump's posts from its content regulation policies and voiced particular concern about the ramifications of this exemption for our political process:

> [W]e have grave concerns that the combination of [Facebook's] decision to exempt politicians from fact-checking and the precedents set by its recent decisions on President Trump's posts, leaves the door open for the platform to be used by other politicians to interfere with voting. If politicians are free to mislead people about official voting methods (by labeling ballots illegal or making other misleading statements that go unchecked, for example) and are allowed to use not-so-subtle dog whistles with impunity to incite violence against groups advocating for racial justice, this does not bode well for the hostile voting environment that can be facilitated by Facebook in the United States. We are concerned that politicians, and any other user for that matter, will capitalize on the policy gaps made apparent by the president's posts and target particular communities to suppress the votes of groups based on their race

triangle, Facebook deactivated those ads because the image was the same symbol used by the Nazis to denote political prisoners in its concentration camps. *Id.* Facebook representatives stated that the ad violated a policy against using a "banned hate group's symbol[s]" outside of a condemnatory context or as an object for discussion. *Id.*

128 Craig Timberg & Elizabeth Dwoskin, *Silicon Valley Is Getting Tougher on Trump and His Supporters over Hate Speech and Disinformation*, WASH. POST (July 10, 2020, 10:53 AM PDT), https://www.washingtonpost.com/technology/2020/07/10/hate-speech-trump-tech/ [https://perma.cc/XE86-62KN].

129 *See* LAURA W. MURPHY, STEPHEN HAYES, ERIC SUBLETT, ALEXA MILTON, TANYA SEHGAL, ZACHARY BEST & MEGAN CACACE, FACEBOOK'S CIVIL RIGHTS AUDIT – FINAL REPORT 5 (2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf [https://perma.cc/6HMU-FYC7].

130 *Id.* at 37-38.

> or other characteristics. . . . [T]his is deeply troublesome as misinformation, sowing racial division and calls for violence near elections can do great damage to our democracy.[131]

The concerns of the auditors turned out to be well-founded, as we now know. Calls for violence near the elections did in fact do great damage to our democracy, and Facebook attempted to belatedly modify its policies to address the grave harms resulting from the dangerous political speech on its platform. In the post-election period, Facebook attempted to slow down the rapidly proliferating election fraud conspiracies and "stop the steal" movement by adding "friction" to engaging with posts, removing the worst violations, and carrying out previously announced election-integrity policies. As I describe in greater detail below, in the period directly after the November election, Facebook initially responded to Trump's false election claims and the burgeoning "stop the steal" movement by adding counter-information to newsfeeds and directing users to official news sources; labeling the worst misinformation cases, including the President's false claims of victory, and banning the largest "stop the steal" Facebook group and related hashtags. Then, in the run up to the insurrection, Facebook extended its temporal ban on political ads, with an exception for the Georgia Senate Runoffs. Finally, after January 6, 2021, Facebook banned Trump indefinitely, announced additional measures to identify and remove content encouraging further incitement of violence, and continued its efforts to combat militarized social movements and to restrict the use of the platform to organize attacks. I detail these efforts below.

First, in the period directly after the election, Facebook turned to counterspeech remedies by adding a notification on top of Facebook and Instagram feeds stating that no winner had been projected for the 2020 Election after former President Trump falsely claimed victory on election night.[132] Facebook also labeled Trump's post claiming victory directing users to news sources Facebook deems credible.[133] These labels were applied to both presidential candidates and other high-profile users in regard to premature declarations of victory or election misinformation.[134] Additionally, Facebook amended its News Feed

---

131 *Id.* at 10.

132 *See* Facebook Newsroom (@fbnewsroom), TWITTER (Nov. 4, 2020, 12:00 AM), https://twitter.com/fbnewsroom/status/1323897798421442566 [https://perma.cc/7ZF7-GVFB].

133 *See id.*

134 *See* Facebook Newsroom (@fbnewsroom), TWITTER (Nov. 4, 2020, 1:18 PM), https://twitter.com/fbnewsroom/status/1324098616911343617 [https://perma.cc/9P3S-

algorithm to prioritize authoritative news sources like NPR, CNN, and *The New York Times* over overtly partisan outlets like Breitbart or Occupy Democrats based on "news ecosystem quality" scores ("NEQ").[135] In addition, forty-eight hours after the election and about twenty-four hours before CNN called the 2020 Election for President Biden, Facebook banned a group of over 350,000 members, many of whom claimed that Democrats were "stealing" the election and some who called for violence.[136] This group was just one of many smaller groups fostering the same sentiments and promoting similar activities.[137] However, this particular group that was shut down had amassed 320,000 of its followers within twenty-two hours of its creation the day after the election.[138] Groups like Women For America First (100,000 followers within a few hours), run by former Georgia congressional candidate Amy Kremer, urged users to join the "Stop the Steal" group. A common goal of these groups was to promote protest efforts in swing states like Pennsylvania and Arizona to disrupt the ongoing vote-counting.[139] In a statement, Facebook spokesperson Andy Stone cited the organized effort around delegitimizing the election and calls for violence in the group "during this period of heightened tension" for its decision to ban the group. Additionally, Facebook stated it would suppress the distribution of election-related livestreams and content related to "stop the steal" efforts — including through banning related hashtags.[140] However, numerous Facebook events for "stop the

XAMQ]; *see also* Rachel Kraus, *Facebook Labeled 180 Million Posts as 'False' Since March. Election Misinformation Spread Anyway*, MASHABLE (Nov. 19, 2020), https://mashable.com/article/facebook-labels-180-million-posts-false/ [https://perma.cc/Z8TD-QQK2].

135 Kevin Roose, Mike Isaac & Sheera Frenkel, *Facebook Struggles to Balance Civility and Growth*, N.Y. TIMES (Nov. 24, 2020), https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html [https://perma.cc/3B2Q-CRWD].

136 Barbara Ortutay & David Klepper, *Facebook Bans Big 'Stop the Steal' Group for Sowing Violence*, AP NEWS (Nov. 5, 2020), https://apnews.com/article/election-2020-donald-trump-misinformation-violence-elections-d5c9bd5fe6a799fd627c50521b6cbb36 [https://perma.cc/64BF-XY37].

137 *Id.*

138 Sheera Frenkel, *The Rise and Fall of the 'Stop the Steal' Facebook Group*, N.Y. TIMES (Nov. 5, 2020), https://www.nytimes.com/2020/11/05/technology/stop-the-steal-facebook-group.html [https://perma.cc/8Y9W-PT89].

139 Tony Romm, Isaac Stanley-Becker & Elizabeth Dwoskin, *Facebook Bans 'STOP THE STEAL' Group Trump Allies Were Using to Organize Protests Against Vote Counting*, WASH. POST (Nov. 5, 2020, 7:01 PM), https://www.washingtonpost.com/technology/2020/11/05/facebook-trump-protests/ [https://perma.cc/YGN3-LW35].

140 *Id.* (noting that Facebook spokesperson Andy Stone called these steps "exceptional measures that we are taking during this period of heightened tension. The

steal" protest events remained active and several "stop the steal" videos had already gone viral. Dozens of smaller "stop the steal" groups began appearing after the "flagship" group had been removed.[141] In addition, posts encouraging members of "stop the steal" groups to visit StolenElection.us, which directed users to join a mailing list, had already been circulated.[142]

In addition, Facebook initially stopped running all "social issue, electoral or political ads" at noon on Wednesday, November 4, 2020.[143] Responding to public criticism, Facebook made an exception to its political ad ban "with the purpose of reaching voters in Georgia about Georgia's runoff elections" starting December 16, 2020.[144] Advertisers directly involved with the elections (campaigns, local election officials, and official political parties) were prioritized while ads targeting locations outside of Georgia and debunked by third-party fact-checkers were prohibited. This was lifted on January 5, 2021 after the elections were completed.[145] On November 11, Facebook announced an extension of the "pause" to last another month, "though there may be an opportunity to resume these ads sooner."[146] It linked to Facebook executive Rob Leathern's tweets for more information.[147] The platform's extension of its ad ban was likely intended to prevent President Trump and his allies from using paid advertising to promote their baseless election fraud claims and attempts to dispute the results.[148]

group was organized around the delegitimization of the election process, and we saw worrying calls for violence from some members of the group").

141 Frenkel, *supra* note 138.

142 Makena Kelly, *Facebook Shuts Down Huge 'Stop the Steal' Group*, VERGE (Nov. 5, 2020, 3:19 PM), https://www.theverge.com/2020/11/5/21551551/facebook-stop-the-steal-group-misinformation-election-2020 [https://perma.cc/732L-GUZE].

143 *U.S. Reminders for When the Polls Close*, FACEBOOK FOR GOV'T, POL. & ADVOC. (Nov. 2, 2020), https://www.facebook.com/gpa/blog/reminders-for-when-the-polls-close [https://perma.cc/4983-FQUS].

144 Sarah Schiff, *An Update on the Georgia Runoff Elections*, FACEBOOK (Dec. 15, 2020), https://about.fb.com/news/2020/12/update-on-the-georgia-runoff-elections/ [https://perma.cc/KX8M-EF6T].

145 *Id.*

146 *Id.*; *5 Things to Remember About Political and Issue Advertising Around the US 2020 Election*, FACEBOOK FOR BUS. (Nov. 11, 2020, 2:45 PM), https://www.facebook.com/business/news/facebook-ads-restriction-2020-us-election [https://perma.cc/8K4X-7GXN].

147 *Id.*; *see* Rob Leathern (@robleathern), TWITTER (Nov. 11, 2020, 1:37 PM), https://twitter.com/robleathern/status/1326640175724847105 [https://perma.cc/49PK-3RDV].

148 Nick Statt, *Facebook Extends Political Ad Ban Another Month as Trump Refuses to Concede*, VERGE (Nov. 11, 2020, 1:25 PM), https://www.theverge.com/2020/11/11/21560969/facebook-political-ad-ban-extended-trump-2020-election-concede-misinformation [https://perma.cc/G3EE-ML2N].

Later, after Twitter took steps to restrict Trump's access in the wake of the January 6 siege on the Capitol, Facebook announced that Trump wouldn't be able to post for twenty-four hours. On January 7, 2021, Facebook announced that it would be extending the block on his accounts indefinitely, for a minimum of two weeks.[149] Mark Zuckerberg personally posted a statement on the decision on his own Facebook page.[150] After placing labels on the posts, Facebook and Instagram eventually removed the video of Mr. Trump condoning the violence and continuing to spread election falsities.[151] On January 21, 2021, Facebook asked the recently constituted Facebook Oversight Board to rule on Facebook's decision to suspend Trump from its platform.[152] The Oversight Board held in May 2021 that, given the seriousness of Trump's violations of Facebook's Community Standards and the ongoing risk of violence they presented, Facebook was justified in suspending Mr. Trump's accounts on January 6, but that it was not appropriate for Facebook to impose an unprecedented indefinite suspension, in the absence of any clear, regular, published procedure for imposing such an indefinite suspension.[153]

In Facebook's January 2021 statement announcing the restrictions on Trump, the platform announced that it would identify and remove content containing, among other content:

- "Praise and support of the storming of the US Capitol . . .

[149] Guy Rosen, *Our Response to the Violence in Washington*, FACEBOOK (Jan. 6, 2021), https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/ [https://perma.cc/42LZ-YC3J] [hereinafter *Our Response*].

[150] Mark Zuckerberg, FACEBOOK (Jan. 7, 2021, 10:47 AM), https://www.facebook.com/zuck/posts/10112681480907401 [https://perma.cc/74L3-3CSL].

[151] Makena Kelly, *Facebook Declares 'Emergency Situation' and Removes Trump Video*, VERGE (Jan. 6, 2021, 5:58 PM), https://www.theverge.com/2021/1/6/22217788/facebook-remove-trump-video-emergency-situation-mob-violence [https://perma.cc/H4WC-ZACG].

[152] Nick Clegg, *Referring Former President Trump's Suspension from Facebook to the Oversight Board*, FACEBOOK (Jan. 21, 2021), https://about.fb.com/news/2021/01/referring-trump-suspension-to-oversight-board/ [https://perma.cc/CSV8-V72P].

[153] Brian Fung, *Facebook's Oversight Board Will Decide Whether Trump Should Be Banned*, CNN BUS. (Jan. 21, 2021), https://www.cnn.com/2021/01/21/tech/facebook-trump-oversight-board/index.html [https://perma.cc/8NU4-HNHN]; *Oversight Board Upholds Former President Trump's Suspension, Finds Facebook Failed to Impose Proper Penalty*, OVERSIGHT BD. (May 2021), https://oversightboard.com/news/226612455899839-oversight-board-upholds-former-president-trump-s-suspension-finds-facebook-failed-to-impose-proper-penalty/ [https://perma.cc/35XU-FAHH].

- Incitement or encouragement of the events at the Capitol, including videos and photos from the protestors . . . [or]
- Attempts to restage violence."[154]

Facebook also announced the following additional measures:

- An update to the text of its labels for posts containing misinformation to read: "Joe Biden has been elected President with results that were certified by all 50 states."
- A continuation of its ban on militarized social movements and QAnon related content — citing its removal of 600 "militarized social movements" from the platform.
- A continuation of pre-existing emergency measures and the implementation of additional ones, including using AI to demote content that likely violates its policies.[155]

In summary, Facebook has undertaken aggressive counterspeech interventions in the form of labeling and commissioning authoritative responses to speech to combat publicly available misinformation on its platform, but, prior to the unprecedented events of the January 6 insurrection, Facebook had excluded politicians' speech from such interventions. In response to the events surrounding the January 6 insurrection, Facebook, following Twitter's lead, undertook unprecedented measures to ban President Trump from its platform in response to Trump's role in inciting the insurrection, as well as to restrict certain types of similar harmful content.

C. Effectiveness of Counterspeech Efforts

The platforms' efforts to engage in forms of counterspeech to combat misinformation appear to have been moderately effective. According to one recent study, social media users were about fifty percent less likely to share false stories if the stories had been labeled as false. When no labels were used at all, participants considered sharing about thirty percent of false stories in the sample, but that figure dropped to about sixteen percent of false stories that had a label attached.[156] In addition,

[154] Rosen, *Our Response*, *supra* note 149.

[155] *Id.*

[156] Peter Dizikes, *The Catch to Putting Warning Labels on Fake News*, MIT NEWS (Mar. 2, 2020), http://news.mit.edu/2020/warning-labels-fake-news-trustworthy-0303 [https://perma.cc/2T78-257E].

the labeling of posts as false led to improved accuracy in social media users' beliefs. Researchers found, in an exhaustive series of surveys across more than 10,000 participants on a wide range of topics, that sixty percent of respondents gave accurate answers when presented with counterspeech in the form of a fact-check/correction, while only thirty-two percent expressed accurate beliefs when they were not presented with such a fact-check/correction.[157] In addition, Hunt Allcott and his co-authors report in their 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."[158] The authors of the study find that user interaction with known false news sites has declined by fifty percent since the 2016 election.[159] 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 [2016] election'[160]] may have had a meaningful impact."[161]

IV. REGULATION OF POLITICAL ADVERTISING AND OF MICROTARGETING OF POLITICAL ADS ON SOCIAL MEDIA

A. Introduction

One of the gravest problems brought about by social media for the marketplace of ideas is the facilitation of filter bubbles, in which

---

[157] Lee Drutman, *Fact-Checking Misinformation Can Work. But It Might Not Be Enough.*, FIVETHIRTYEIGHT (June 3, 2020, 1:01 PM), https://fivethirtyeight.com/features/why-twitters-fact-check-of-trump-might-not-be-enough-to-combat-misinformation/ [https://perma.cc/R6KA-BZGM]. The political scientists conducting the surveys, Ethan Porter and Thomas J. Wood, found that the most effective fact-checks shared four characteristics: they were from a highly credible source, they offered a new frame for the issue rather than merely calling the misinformation "wrong," they didn't directly challenge a worldview or identity, and they happened before a false narrative could gain traction. *Id.*

[158] Hunt Allcott, Matthew Gentzkow & Chuan Yu, *Trends in the Diffusion of Misinformation on Social Media* 1 (Stanford Inst. for Econ. Policy Research, Working Paper No. 18-029, 2018), https://web.stanford.edu/~gentzkow/research/fake-news-trends.pdf [https://perma.cc/42VB-TXJX].

[159] *Id.* at 5.

[160] *Id.* at 3, 6.

[161] *Id.* at 3.

members of the public are able to insulate themselves from diverse and antagonistic viewpoints and from effective counterspeech. The fractionation and self-isolation of members of the citizenry pose grave problems for our information ecosystem and for our democracy, especially given the ability to microtarget advertisements — and especially political advertisements — via social media platforms, as I describe below.

Microtargeting of advertisements on social media platforms is the practice that generally allows advertisers to limit their messaging to narrow slices or subsets of individuals by exploiting the vast trove of social data about individuals' online behavior and preferences that has been collected by social media platforms.[162] Microtargeting of advertisements in general stands in sharp contrast to the broadcasting of ads in legacy media like major metropolitan newspapers, radio and television, through which advertisers provide content to a broad audience (e.g., to all readers of *The Washington Post*). In contrast, microtargeting on social media delivers ad content to very specific subgroups (e.g., readers who shop at Whole Foods who are between the ages of twenty-five and forty-nine, and who have watched a certain video on YouTube) or even to specific, listed individuals (by using tools such as Facebook's Custom Audiences).[163] The practice of microtargeting employs and capitalizes on the social data — such as an individual's likes, dislikes, interests, preferences, behaviors and viewing and purchasing habits — collected by social media platforms about their users and made available to advertisers to enable advertisers to segment individuals into small groups so as to more accurately and narrowly target advertising to them.[164] Facebook, for example, reportedly tracks a list of over 1,100 attributes on each of its users spanning users' demographic, behavioral, and interest categories.[165]

---

[162] *Microtargeting*, INFO. COMMISSIONER'S OFF., https://ico.org.uk/your-data-matters/be-data-aware/social-media-privacy-settings/microtargeting/ (last visited Oct. 16, 2020) [https://perma.cc/H7N9-NZ7E].

[163] *See* Dipayan Ghosh, *What Is Microtargeting and What Is It Doing in Our Politics?*, MOZILLA: INTERNET CITIZEN (Oct. 4, 2018), https://blog.mozilla.org/internetcitizen/2018/10/04/microtargeting-dipayan-ghosh/ [https://perma.cc/82AN-DBVU].

[164] *Id.*

[165] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau & Alan Mislove, *Potential for Discrimination in Online Targeted Advertising*, 81 PROC. MACHINE LEARNING RES. 1, 7 (2018) ("For each user in the US, Facebook tracks a list of over 1,100 binary attributes spanning demographic, behavioral and interest categories that we refer to as *curated attributes*. Additionally, Facebook tracks users' interests in entities such as

The practice of microtargeting enables advertisers to capitalize on the comprehensive social data about individuals collected by social media platforms. This social data is then used to design and disseminate content that advertisers predict will be the most effective and relevant with respect to the targeted segment of individuals. For example, an advertiser might limit the scope of an ad's distribution to "single men between 25 and 35 who live in apartments and 'like' the Washington Nationals."[166] While businesses derive certain benefits from the microtargeting of ads in nonpolitical contexts, microtargeting of ads in the political context can pose serious problems for the democratic process and for the marketplace of ideas model that underlies our First Amendment model of freedom of speech.[167] Unlike political advertising on mass media like broadcast television or radio — in which large national or regional audiences are exposed to the same political advertisement — by employing narrowly cast microtargeted ads on social media, a political advertiser can craft a specific ad to a much narrower intended audience, and to *only* that specific audience, thereby preventing others from accessing and scrutinizing the content of the ad.

As described by Facebook's former Chief Security Officer Alex Stamos, the chief benefit of political microtargeting is that it allows political advertisers to deploy "messages that are extremely finely targeted to a very small number of people."[168] By microtargeting political ads, a campaign can make different, and even contradictory, appeals to voters in Michigan and to voters in New York or Atlanta. As such, extensively deployed microtargeting of political ads — which is by definition immune from the check of broad public scrutiny — increases the possibility that a politician might lie with impunity. As Stamos explains, "[i]f you allow people to show an ad to just 100 folks, and then you run tens of thousands of ads, then it makes it extremely

websites, apps, and services as well as topics ranging from food preferences (e.g., pizza) to niche interests (e.g., space exploration)." (emphasis in original)).

166 Ellen L. Weintraub, Opinion, *Don't Abolish Political Ads on Social Media. Stop Microtargeting.*, WASH. POST (Nov. 1, 2019, 6:51 PM), https://www.washingtonpost.com/opinions/2019/11/01/dont-abolish-political-ads-social-media-stop-microtargeting/ [https://perma.cc/C9L9-QZ4V].

167 *See generally* Nunziato, *The Marketplace of Ideas Online*, *supra* note 5 (explaining the constitutional relevance of the marketplace of ideas model and illustrating the ways in which regulation of online speech and microtargeting impact it).

168 Peter Kafka, *Facebook's Political Ad Problem, Explained by an Expert*, VOX (Dec. 10, 2019, 8:00 AM), https://www.vox.com/recode/2019/12/10/20996869/facebook-political-ads-targeting-alex-stamos-interview-open-sourced [https://perma.cc/6WHK-9P8T].

difficult for your political opponent and the print media to call you out."[169]

Microtargeting of political ads also exacerbates problems of balkanization, in which the messages that individuals receive are so disparate as to dissolve the larger communities of interest that otherwise ostensibly bind the country as a nation.[170] A recent study in fact showed that the very mechanism of Facebook's ad delivery increases partisanship.[171] The authors of the study isolated the role that Facebook's perception of an ad's content plays in determining the audience that receives it by creating a generic, non-partisan ad with a call to register to vote that linked to a generic domain.[172] The authors then "configured [their] web server to deliver a different response for requests for these pages based on the IP address of the requestor."[173] If the requestor were identified as Facebook, it would be served "a copy of the HTML from the official Trump campaign website, the official Sanders campaign website, or a generic voting information website."[174] All other requestors were simply "redirected to the generic voting information website."[175] The ads therefore appeared identical to users, but misled Facebook's algorithm to associate them with different political content.[176] The authors found that even after selecting a target audience, Facebook will prefer delivering the ad to those it predicts will identify with its message.[177] The authors conclude that, "[c]ounterintuitively, advertisers who target broad audiences may end

169 *Id.*

170 *See* Craig Timberg, *Critics Say Facebook's Powerful Ad Tools May Imperil Democracy. But Politicians Love Them.*, WASH. POST (Dec. 9, 2019, 9:00 AM), https://www.washingtonpost.com/technology/2019/12/09/critics-say-facebooks-powerful-ad-tools-may-imperil-democracy-politicians-love-them/ [https://perma.cc/848J-BL8L]; *see also* Isaac Stanley-Becker, *Facebook's Ad Tools Subsidize Partisanship, Research Shows. And Campaigns May Not Even Know It.*, WASH. POST (Dec. 10, 2019, 8:00 AM), https://www.washingtonpost.com/technology/2019/12/10/facebooks-ad-delivery-system-drives-partisanship-even-if-campaigns-dont-want-it-new-research-shows/ [https://perma.cc/6HHE-RXDU] (explaining that serving "users with information that aligns with their existing worldview . . . 'fragments political discourse'").

171 *See* MUHAMMAD ALI, PIOTR SAPIEZYNSKI, ALEKSANDRA KOROLOVA, ALAN MISLOVE & AARON RIEKE, AD DELIVERY ALGORITHMS: THE HIDDEN ARBITERS OF POLITICAL MESSAGING 13 (2019), https://arxiv.org/pdf/1912.04255.pdf [https://perma.cc/SX3B-UFZE].

172 *Id.* at 7, 13.

173 *Id.* at 7.

174 *Id.*

175 *Id.*

176 *Id.* at 7-8.

177 *See id.* at 9-10.

up ceding [to] platforms even more influence over which users ultimately see which ads."[178] Beyond the "*ad creation and targeting* phase, where the advertiser selects their desired audience," the actual delivery of the ad further discriminates among possible recipients.[179] The selection is "rooted in the desire to show *relevant* ads to users" and, the study notes, "can lead to dramatic skew in delivery along gender and racial lines, even when the advertiser aims to reach gender and race-balanced audiences."[180]

The Internet Research Agency — the notorious agent of Russian disinformation during the 2016 election cycle — was able to spend pennies on the dollar (or ruble) compared to U.S. presidential campaigns by deploying powerful microtargeted political ads on social media. With its use of microtargeted political ads, the Agency was able to powerfully leverage its influence to interfere with U.S. elections. While the Trump and Clinton campaigns spent a combined $81 million on pre-election Facebook ads,[181] for example, the IRA was able to sow tremendous discord by spending only $46,000.[182] This miniscule amount of spending took advantage of the powerful ability to target custom audiences by inferring interests from social media users' social data. The Internet Research Agency used the microtargeting tools developed by leading technology companies — including Facebook's advertising customization tools — to target specific audiences that they believed would be particularly susceptible to false and misleading election-related information. In particular, Russian operatives used

[178] *Id.* at 1.

[179] *Id.* (emphasis in original).

[180] *Id.* at 1-2 (emphasis in original).

[181] Josh Constine, *Trump and Clinton Spent $81M on US Election Facebook Ads, Russian Agency $46k*, TECHCRUNCH (Nov. 1, 2017, 8:38 AM PDT), https://techcrunch.com/2017/11/01/russian-facebook-ad-spend [https://perma.cc/K43L-7B8X].

[182] *See Social Media Influence in the 2016 U.S. Election: Hearing Before the Senate Select Comm. on Intel.*, 115th Cong. 76 (2017), https://www.intelligence.senate.gov/hearings/open-hearing-social-media-influence-2016-us-elections# [https://perma.cc/25UZ-GDFE] (testimony of Colin Stretch, General Counsel, Facebook). For Facebook VP of Advertising Rob Goldman's ham-fisted reaction to Russia's ad spending, see Kevin Roose, *On Russia, Facebook Sends a Message It Wishes It Hadn't*, N.Y. TIMES (Feb. 19, 2018), https://www.nytimes.com/2018/02/19/technology/russia-facebook-trump.html [https://perma.cc/6LNY-QJLK]; Nicholas Thompson, *A Facebook Executive Apologizes to His Company — and to Robert Mueller*, WIRED (Feb. 19, 2018, 11:47 PM), https://www.wired.com/story/facebook-executive-rob-goldman-apologizes-to-company-and-robert-mueller/ [https://perma.cc/PN5F-5Z5Q].

Facebook's Custom Audiences[183] tool to display specific ads and messages to voters who had visited the operatives' fake social media sites — and used this microtargeting technique to sew division among voters — specifically to suppress Black voter turnout.[184] Facebook's Custom Audiences tool allows advertisers, including, in this case, the Russian operatives, to input into Facebook's system a specific list of users they wish to target. While such technological tools have long been used by corporate America to deliver advertising to target audiences, Facebook and other social media platforms were taken by surprise by the use of such tools for purposes of interference in the U.S. elections. As *The Washington Post* explains: Russian operatives' microtargeted political ads

> focused on such hot-button issues as illegal immigration, African American political activism and the rising prominence of Muslims in the United States. The Russian operatives then used a Facebook "retargeting" tool, called Custom Audiences, to send specific ads and messages to voters who had visited those sites. . . . One such ad featured photographs of an armed Black woman "dry firing" a rifle — pulling the trigger of the weapon without a bullet in the chamber. . . . Investigators believe the advertisement may have been designed to encourage African American militancy and, at the same time, to stoke fears within white communities . . . .[185]

Russian operatives used other Facebook tools in addition to Custom Audiences to target groups by demographics, geography, gender, and interests. As Clinton Watts, a fellow at the Foreign Policy Research Institute, explains, "This means that any American who knowingly or unknowingly clicked on a Russian news site may have been targeted through Facebook's advertising systems to become an agent of influence — a potentially sympathetic American who could spread Russian propaganda with other Americans."[186] Accordingly, "[e]very successful

---

[183] *About Website Custom Audiences*, FACEBOOK FOR BUS., https://www.facebook.com/business/help/610516375684216 (last updated Sept. 27, 2019) [https://perma.cc/J2UZ-7E2V].

[184] Spencer Overton, *State Power to Regulate Social Media Companies to Prevent Voter Suppression*, 53 UC DAVIS L. REV. 1793, 1795-96 (2020).

[185] Elizabeth Dwoskin, Craig Timberg & Adam Entous, *Russians Took a Page from Corporate America by Using Facebook Tool to ID and Influence Voters*, WASH. POST (Oct. 2, 2017), https://www.washingtonpost.com/business/economy/russians-took-a-page-from-corporate-america-by-using-facebook-tool-to-id-and-influence-voters/2017/10/02/681e40d8-a7c5-11e7-850e-2bdd1236be5d_story.html [https://perma.cc/6QKB-ARZF].

[186] *Id.*

click [provides the Russian operatives with] more data that they can use to retarget. . . . [thereby speeding up] the influence dramatically."[187] Targeted Facebook users were then shown ads featuring divisive topics that the Russians wanted to promote in their Facebook news feeds, which displayed the ads alongside messages from friends and family members.

The Russian Internet Research Agency notoriously used Facebook's complex ad targeted tools to microtarget political ads to African Americans in order to suppress the Black vote in the 2016 election.[188] African American audiences accounted for over thirty-eight percent of U.S.-focused ads purchased by the Internet Research Agency, which created social media accounts that falsely claimed they were African American-operated and urged African Americans to "boycott the election." As Professor Spencer Overton explains:

> Facebook's "Ad Manager" allows an advertiser to select, from a series of dropdowns, 52,000 targeting attributes, including demographics/ethnic affinity (e.g., African American), issue interests (e.g., "Malcolm X" or the "Civil Rights Movement"), and Facebook engagement (e.g., liked a particular post). . . . Facebook develops these profiles by collecting vast amounts of data on its two billion users — including zip codes, posts, comments, likes, clicks, and other information — and by utilizing predictive modeling techniques to make inferences.[189]

In short, using Facebook's powerful microtargeting tools, Russian operatives were able to target African-American members of our electorate, sow division, and — among other problems — suppress the Black vote.

The Internet Research Agency was not alone in its masterful deployment of microtargeted political ads in the 2016 presidential election. The Trump Campaign, for example, also targeted Black Americans in specific neighborhoods in an effort to decrease voter participation.[190] The benefit, as then-Trump digital media director Brad Parscale described it, was that "only the people we want to see it, see it."[191] Parscale claimed that the use of microtargeted political ads on

---

[187] *Id.*

[188] *See* Overton, *supra* note 184, at 1815.

[189] *Id.*

[190] Joshua Green & Sasha Issenberg, *Inside the Trump Bunker, with Days to Go*, BLOOMBERG (Oct. 27, 2016, 3:00 AM PDT), https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go [https://perma.cc/NU3F-ABS2].

[191] *Id.*

Facebook and Twitter enabled the Trump Campaign to be one hundred to two hundred times more effective in targeting members of the electorate than the Hillary for President Campaign.[192] Whether or not Parscale's particular claim is true, research shows that political microtargeting indeed had a significant effect "in persuading undecided voters to support Mr. Trump, and in persuading Republic supporters to turn out on polling day."[193] Specifically, researchers found that "targeted Facebook campaigning increased the probability that a previously non-aligned voter would vote for Donald Trump, by at least five percent" if they were a regular Facebook user.[194]

The microtargeting of political ads, compared to the dissemination of political ads via traditional media outlets, is problematic for a number of reasons from the perspective of First Amendment values, and this is not even considering the problems caused by the weaponization of microtargeting by Russian operatives interfering in our elections, sowing division, and suppressing the Black vote. First, political ads disseminated via traditional media are subject to broad exposure and broad public scrutiny — which are necessary for the truth-facilitating features of the marketplace of ideas mechanisms to function. Microtargeted ads, on the other hand — which are essentially the "online equivalent of whispering millions of different messages into zillions of different ears for maximum effect and with minimum scrutiny"[195] — are not similarly subject to broad exposure or broad public scrutiny. Second, and relatedly, microtargeted ads on social media are more likely to be susceptible to the spread of misinformation. As politics and technology expert Dipayan Ghosh explains:

> [Microtargeting of political ads facilitates] 'organic' shares and reshares of content pushed by unpaid users who appreciate what they see . . . and wish to spread it around their networks. This results in free content consumption for the political campaign. . . . [and this] viral spread of 'unpaid' or 'organic'

[192] Brad Parscale (@parscale), TWITTER (Feb. 24, 2018, 1:46 PM), https://twitter.com/parscale/status/967516077956755457 [https://perma.cc/3JH4-DQFD].

[193] Univ. of Warwick, *Targeted Facebook Ads Shown to Be Highly Effective in the 2016 US Presidential Election*, PHYS.ORG (Oct. 25, 2018), https://phys.org/news/2018-10-facebook-ads-shown-highly-effective.html [https://perma.cc/54S3-XQPE].

[194] Federica Liberini, Michela Redoano, Antonio Russo, Angel Cuevas & Ruben Cuevas, *Politics in the Facebook Era: Evidence from the 2016 US Presidential Elections* 5 (Ctr. for Competitive Advantage in the Glob. Econ., Working Paper No. 389, 2018).

[195] Kara Swisher, *Google Changed Its Political Ad Policy. Will Facebook Be Next?*, N.Y. TIMES (Nov. 22, 2019), https://www.nytimes.com/2019/11/22/opinion/google-political-ads.html [https://perma.cc/P832-RLD5].

> content . . . further encourages the success of misinformation campaigns.[196]

In short, the microtargeting of political ads disseminated via social media, and especially via Facebook, is especially pernicious because it is not subject to meaningful widespread public scrutiny — and because false claims in such political ads are likely to be spread farther, faster, deeper, and more broadly than true claims in political ads.[197]

B. Facebook

Facebook has been the primary social media platform facilitating the microtargeting of political ads to members of the public, allowing political advertisers to have access to the vast trove of social data that it collects on its users, to serve up ads to users with great precision and with no public scrutiny. The company has also been unwilling to prohibit the microtargeting of political ads on its platform, despite many calls for it to do so, including by the chair of the Federal Elections Commission.[198] The company's continued allowance of microtargeting of political ads on its platform has posed grave problems for the marketplace of ideas and the counterspeech mechanism. Although Facebook in late 2019 was reportedly considering increasing the minimum number of people who can be targeted in political ads on its platform from 100 to a few thousand, as of this date, Facebook has not made any changes to its policy allowing for the microtargeting of political ads.[199]

---

[196] Ghosh, *supra* note 163.

[197] According to a recent study published in *Science*, false news — and in particular, false political news — spreads more quickly than the truth, with the top one percent of false news cascades diffused to between 1,000 and 100,000 people (whereas the truth rarely diffused to more than 1,000 people) and with false news diffusing faster than the truth. The authors of the study investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017; this included approximately 126,000 stories tweeted by approximately three million people more than 4.5 million times. They observed that "[f]alsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news" than for false news concerning other subjects, such as "natural disasters, science, urban legends, or financial information." Soroush Vosoughi, Deb Roy & Sinan Aral, *The Spread of True and False News Online*, 359 SCIENCE 1146, 1146, 1148 (2018).

[198] *See* Rob Leathern, *Expanded Transparency and More Controls for Political Ads*, FACEBOOK NEWSROOM (Sept. 14, 2020), https://about.fb.com/news/2020/01/political-ads/ [https://perma.cc/U4TL-GKHW].

[199] *See* Emily Glazer, *Facebook Weighs Steps to Curb Narrowly Targeted Political Ads*, WALL ST. J. (Nov. 21, 2019, 8:13 PM ET), https://www.wsj.com/articles/facebook-

What Facebook has done is to increase transparency and disclosure requirements for political advertisements, so that such ad practices can (theoretically) be analyzed and scrutinized.[200] Facebook also recently implemented a Political Advertising Policy that mandates labeling, disclosure, and transparency requirements on political ads. Under this Policy, every election-related and issue advertisement made available on Facebook to users in the United States must 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.[201] Second, under the Policy, Facebook collects and maintains a publicly available archive of all political advertisements made available on its platform as part of its Ad Library. The Facebook Ad Library provides information regarding the budget associated with each ad and how many people saw it, including their age, location, and gender,[202] as can be seen in the example below, in which a group called "Inner Americans" spend under $100 to largely target men over age forty-five with an ad entitled "Nancy Pelosi Goes On Unhinged Rant,

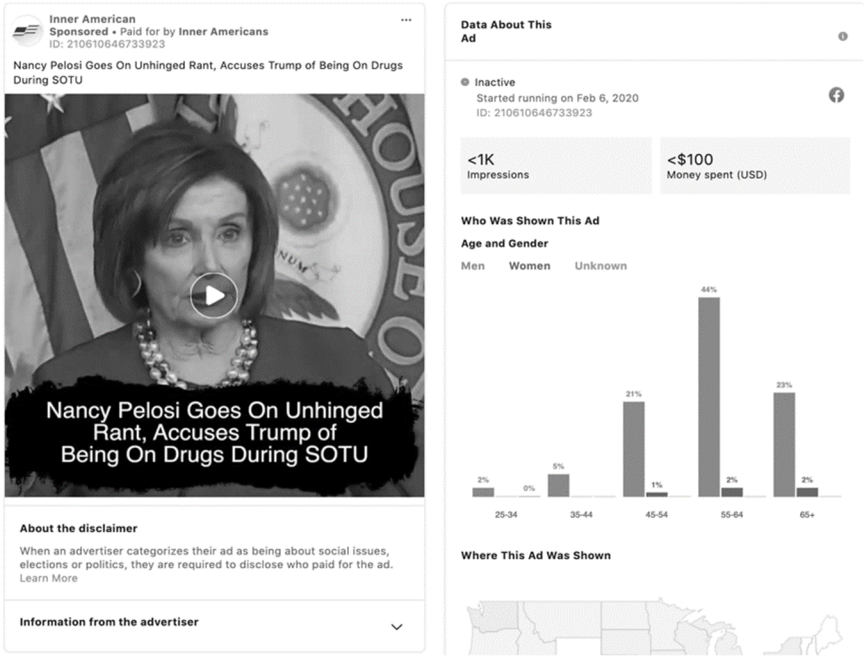
discussing-potential-changes-to-political-ad-policy-11574352887 [https://perma.cc/YX5C-X4XH]; Associated Press, *Facebook Refuses to Restrict Untruthful Political Ads and Micro-Targeting*, GUARDIAN (Jan. 9, 2020, 9:13 AM EST), https://www.theguardian.com/technology/2020/jan/09/facebook-political-ads-micro-targeting-us-election [https://perma.cc/D3UD-BGZP]. The company has, however, adopted a policy that will allow users to see fewer political ads. Emily Birnbaum, *Facebook Will Still Allow Misinformation, Microtargeting Under New Ad Rules*, HILL (Jan. 9, 2020, 8:25 AM EST), https://thehill.com/policy/technology/477486-facebook-will-still-allow-misinformation-micro-targeting-under-new-ad-rules [https://perma.cc/AP3L-BKD7].

200 Facebook recently updated its Ad Library's functionality in an effort to increase transparency and provide enhanced tools to researchers, advocates, and the public generally — including by permitting users to search for and filter ads based on the estimated audience size — which enables researchers to identify and study micro-targeted ads. MURPHY ET AL., *supra* note 129, at 35-37.

201 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/ [https://perma.cc/EJQ8-9MXK]. In addition, under the Policy, Facebook prohibits foreign entities from purchasing political ads directed at U.S. audiences. *See Get Authorized to Run Ads About Social Issues, Elections or Politics*, FACEBOOK FOR BUS., https://www.facebook.com/business/help/208949576550051 (last visited Oct. 16, 2020) [https://perma.cc/4RML-9N8N].

202 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/ [https://perma.cc/EG78-WND2]; *see also* MURPHY ET AL., *supra* note 129, at 36 ("Since 2018, Facebook has maintained a library of ads about social issues, elections or politics that ran on the platform. These ads are either classified as being about social issues, elections or politics or the advertisers self-declare that the ads require a 'Paid for by' disclaimer.").

Accuses Trump of Being on Drugs During SOTU [State of the Union Address]."



[Image taken from: Facebook Ad Library, https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=US&id=210610646733923&view_all_page_id=107371014057680 (click "See Ad Details") (last accessed Oct. 17, 2020) [https://perma.cc/7TSF-6QV8].]

Given the extensive use of Facebook's platform for political speech by politicians and for the dissemination of political ads, and given the ability to engage in the practice of microtargeting of political ads via the platform, the company's decision not to subject such speech to external fact-checking and its decision to exempt such speech from generally-applicable counterspeech measures are ill-founded. Facebook maintains, as mentioned above, that public scrutiny of the political speech on its platform is an adequate substitute for the fact-checking and counterspeech processes that all other public posts on Facebook are now subject to. Given the ability to microtarget political ads and given the filter bubbles in which many Facebook users exist, Facebook's approach regarding political ads is ill-founded. Facebook did, however, impose a blackout period prohibiting new advertisements about social issues, elections, and politics in the week leading up to the 2020

presidential election[203] — and continued this blackout period indefinitely after the election in light of President Trump's refusal to concede. Apparently, Facebook did so because it was concerned that the corrective of counterspeech would be ineffective within this crucial and limited time period to impose meaningful checks on false or misleading content in such advertisements.[204]

C. Twitter

Twitter has taken a different approach to the issues posed by microtargeting of political ads. The company has taken the most aggressive stance of the three major platforms by banning "political ads" altogether. This prohibition applies only to the paid promotion of political content. That is, a politician (or others) may still tweet regarding the politician's qualifications for office and reasons to support him or her, but may not make such an appeal the subject of paid advertising on Twitter; however, this ban does not affect "organic" content or messages from politicians that are shared or retweeted. Twitter CEO Jack Dorsey explained that this restriction was not intended to be a limitation on free expression, but rather one restricting politicians from "paying for reach."[205] Twitter announced its political advertising ban in November 2019, one year before the 2020 presidential election.[206] The policy defines political content as that which "references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome."[207] Ads that reference the above — including by "appeals for votes, solicitations of financial support, and advocacy for or against any of the above-listed types of political content" — are prohibited.[208] Twitter exempts "cause-based"

---

[203] Mark Zuckerberg, FACEBOOK (Sept. 3, 2020), https://www.facebook.com/zuck/posts/10112270823363411 [https://perma.cc/CAG6-YXVJ].

[204] *Id.*

[205] Jack Dorsey (@jack), TWITTER (Oct. 30, 2019, 4:05 PM), https://twitter.com/jack/status/1189634377057067008 [https://perma.cc/863L-BF54].

[206] *See Political Content*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html (last visited Mar. 4, 2021) [https://perma.cc/9QK3-E29U].

[207] *Id.*

[208] PACs, SuperPACs, candidates, political parties, and elected or appointed government officials are also banned from advertising on Twitter. *Political Content FAQs*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content/political-content-faqs.html (last visited Mar. 4, 2021) [https://perma.cc/83CD-3XGB].

ads — ads that "educate, raise awareness, and/or call for people to take action in connection with civic engagement, economic growth, environmental stewardship, or social equity causes"[209] by groups other than political organizations, candidates, or politicians[210] from its blanket ban on political ads,[211] but provides that cause-based ads may not be microtargeted. [212]

D. Google

Of the three major social media platforms, Google has undertaken the most targeted approach to regulate microtargeting of political advertisements. In November 2019, Google amended its rules to restrict microtargeting so that political advertisers can only target ads based on three characteristics: an individual's age, gender, and general location (defined by postal code).[213] Political advertisers can also use contextual targeting, which enables them to serve users with ads according to the content that users are accessing.[214] Under Google's rules, only the following characteristics may be used to target election ads: geographic location (but not radius around a location), age, gender, and contextual targeting options such as ad placements, topics, keywords against sites, apps, pages, and videos.[215] All other types of targeting are not allowed for use in election ads, including the use of Google's powerful Audience Targeting products, Remarketing, Customer Match, and Geographic

---

[209] *Cause-Based Advertising Policy*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/restricted-content-policies/cause-based-advertising.html (last visited Mar. 4, 2021) [https://perma.cc/UQ3P-KH8E].

[210] Vijaya Gadde (@vijaya), TWITTER (Nov. 15, 2019, 10:30 PM), https://twitter.com/vijaya/status/1195408747926917120 [https://perma.cc/R82T-PALR].

[211] *See Cause-Based Advertising Policy*, *supra* note 209.

[212] *See id.* Cause-based advertisers must also undergo a certification process. *Id.; see also Cause-Based Advertiser Certification*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/ads-content-policies/cause-based-advertising/cause-based-certification.html (last visited Mar. 4, 2021) [https://perma.cc/HT9Y-5B2Q].

[213] Scott Spencer, *An Update on Our Political Ads Policy*, GOOGLE BLOG: KEYWORD (Nov. 20, 2019), https://blog.google/technology/ads/update-our-political-ads-policy/ [https://perma.cc/9R8D-MQNR].

[214] *Id.*

[215] *Political Content*, GOOGLE ADVERT. POLICIES HELP, https://support.google.com/adspolicy/answer/6014595 (last visited Mar. 5, 2021) [https://perma.cc/HVM6-BZC5] [hereinafter *Political Content (Google)*]. Like Facebook, Google has also implemented a host of procedural requirements for political advertisers. Advertisers who wish to purchase and run election ads or use political affiliation in personalized advertising in the United States must meet Google's verification requirements. *See id.*

Radius Targeting.[216] Google's microtargeting policy applies to ads shown to users of Google's search engine and YouTube, as well as display advertisements sold by Google that appear on other websites.[217] Election ads will no longer be allowed to target what are called "affinity audiences" that look like other groups that campaigns might want to target.[218] Further, political campaigns can no longer upload their own lists of people to whom they wish to show ads.[219] In addition, Google will prohibit what is known as "remarketing," the process of serving ads to people who have previously taken an action like visiting a campaign's website.[220] Google's microtargeting policy prevents political advertisers from taking advantage of some of Google's most sophisticated targeting tools, upon which it has built its dominant market position.[221] The most granular of those targeting tools are custom audiences (formerly known as "custom affinity" audiences), an offering that has allowed advertisers to create tailor-made audiences by targeting individual interests and lifestyles as defined by keyword phrases.[222] Google's sophisticated targeting tools also have allowed advertisers to target or exclude according to demographic data such as age, gender, household income, homeownership, and the like.[223] General advertisers may also target

[216] *See generally About Audience Targeting*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/2497941 (last visited Mar. 5, 2021) [https://perma.cc/E79C-MZ2X]; *About Customer Match*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/6379332 (last visited Mar. 5, 2021) [https://perma.cc/QG33-KKME]; *About Remarketing*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/2453998 (last visited Mar. 5, 2021) [https://perma.cc/7TFF-3PGP]; *Target Ads to Geographic Locations*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/1722043 (last visited Mar. 5, 2021) [https://perma.cc/L3SP-XS4H].

[217] Spencer, *supra* note 213.

[218] Jenna Lowenstein (@just_jenna), TWITTER (Nov. 20, 2019, 3:54 PM), https://twitter.com/just_jenna/status/1197302201938567168 [https://perma.cc/YXN9-RHNQ]; *see also Political Content (Google)*, *supra* note 215.

[219] *Political Content (Google)*, *supra* note 215.

[220] *Id.*

[221] Patience Haggin & Kara Dapena, *Google's Ad Dominance Explained in Three Charts*, WALL ST. J. (June 17, 2019, 5:30 AM ET), https://www.wsj.com/articles/why-googles-advertising-dominance-is-drawing-antitrust-scrutiny-11560763800 [https://perma.cc/FHN5-RG6A] ("[Google] has a 37% share of the $130 billion U.S. digital ad market, according to research firm eMarketer . . . .").

[222] *About Custom Audiences*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/9805516?hl=en&ref_topic=3122880 (last visited Mar. 7, 2021) [https://perma.cc/SLB9-XWQS].

[223] *About Demographic Targeting*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/2580383?hl=en&ref_topic=3122881 (last visited Mar. 7, 2021) [https://perma.cc/8EVQ-8MXG].

users who have previously interacted with their site[224] or submit previously collected customer data to re-engage with the same group or expand to similar audiences.[225] These sophisticated targeting tools are now unavailable to political advertisers.[226]

Google's limitations on the microtargeting of political ads constitute an important and powerful measure in addressing the problems that filter bubbles pose to our democratic process and to our marketplace of ideas online. While Twitter arguably went too far in addressing the problem of microtargeted political ads — by banning political ads altogether — and while Facebook has not adopted any measures to address the problems caused by the microtargeting of political ads (problems that are mostly of Facebook's own creation[227]), Google's approach appears to be a measured, effective, and appropriately targeted one.

CONCLUSION

In summary, the counterspeech efforts undertaken by the major social media platforms appear to have been moderately effective in combatting falsehoods, limiting the dissemination of false or misleading content, and bringing about the truth in the online marketplace of ideas. The efforts undertaken by the major social media platforms to engage in counterspeech to combat political and election-related misinformation — by labeling harmful content and developing and referring users to accurate information — and by imposing some restrictions on the microtargeting of political ads is largely consistent with First Amendment values and with the marketplace of ideas theory of the First Amendment, according to which the accepted response to bad speech is not censorship but more (better) speech. In addition, the platforms' efforts in countering such misinformation contributes toward "producing an informed public capable of conducting its own affairs" and facilitating the preconditions necessary for citizens to

---

[224] *Remarketing: Reach People Who Visited Your Site or App*, GOOGLE ADS HELP, https://support.google.com/google-ads/topic/3122874 (last visited Mar. 7, 2021) [https://perma.cc/V2VV-EA7Z].

[225] *About Customer Match*, *supra* note 217.

[226] *Political Content (Google)*, *supra* note 215.

[227] As discussed above in the text accompanying notes 197–98, microtargeting employs psychographic data — likes, dislikes, interests, preferences, behaviors, and viewing and purchasing habits — collected by social media platforms about their users. These platforms then make the data available to advertisers to enable more narrowly targeted advertising. Thus, the problems caused by microtargeting are fairly attributable to the social media platforms themselves.

engage in the task of democratic self-government,[228] which are also foundational goals of our First Amendment jurisprudence.

[228] Red Lion Broad. Co. v. FCC, 395 U.S. 367, 392 (1969).