MISINFORMATION MAYHEM: SOCIAL MEDIA PLATFORMS' EFFORTS TO COMBAT MEDICAL AND POLITICAL MISINFORMATION

Dawn Carla Nunziato[*]

Table of Contents

INTRODUCTION

Social media platforms today are playing an ever-expanding role in shaping the contours of today's information ecosystem.[1] The events of recent months have driven home this development, as the platforms have shouldered the burden and attempted to rise to the challenge of ensuring that the public is informed—and not misinformed—about matters affecting our democratic institutions in the context of our elections, as well as about matters affecting our very health and lives in the context of the pandemic. This Article examines the extensive role social media companies have recently assumed in combatting misinformation and disinformation[2] in the online marketplace of ideas, with an emphasis on their efforts to combat medical misinformation in the context of the COVID-19 pandemic as well as their efforts to combat false political speech in the 2020 election cycle. In the context of medical misinformation surrounding the COVID-19 pandemic, this Article analyzes the extensive measures undertaken by the major social media platforms to combat such misinformation. In the context of misinformation in the political sphere, this Article examines the distinctive problems brought about by the microtargeting of

* William Wallace Kirkpatrick Research Professor and Professor of Law, The George Washington University Law School. I am extremely grateful to Nathaniel Christiansen, Chris Frascella, Conor Kelly, Ken Rodriguez, and Jake Warren for providing excellent research and library assistance in connection with this article, to Lindsay Byers and Ashley Nicole Fox for excellent editorial work, and to Associate Dean for Research and Faculty Development Thomas Colby and Dean Dayna Matthew for academic and financial support of my research.

[1] *See generally* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

[2] In this article, I will use the term "misinformation" to refer to false information regardless of whether the speaker of that information had an "intent to mislead," and I use the term "disinformation" to refer to intentionally false information, where the speaker of that information had an intent to mislead. *See, e.g.*, Valerie Strauss, *Word of the Year: Misinformation. Here's Why.*, WASH. POST (Dec. 10, 2018), https://www.washingtonpost.com/education/2018/12/10/word-year-misinformation-heres-why/.

political speech and by false political ads on social media in recent years, and the measures undertaken by major social media companies to address such problems. In both contexts, this Article examines the extent to which such measures are compatible with First Amendment substantive and procedural values.

Social media platforms are essentially attempting to address today's serious problems alone, in the absence of federal or state regulation or guidance in the United States. Despite the major problems caused by Russian interference in our 2016 elections, the U.S. has failed to enact regulations prohibiting false or misleading political advertising on social media—whether originating from foreign sources or domestic ones—because of First Amendment, legislative, and political impediments to such regulation. Additionally, the federal government has failed miserably in its efforts to combat COVID-19 or the medical misinformation that has contributed to the spread of the virus in the U.S. All of this essentially leaves us (in the United States, at least) solely in the hands, and at the mercy, of the platforms themselves—to regulate our information ecosystem (or not), as they see fit.

The dire problems brought about by medical and political misinformation online in recent months and years have ushered in a sea change in the platforms' attitudes and approaches toward regulating content online. In recent months, for example, Twitter has evolved from being the non-interventionist "free speech wing of the free speech party"[3] to designing and operating an immense operation for regulating speech on its platform, epitomized by its recent removal[4] and labeling[5] of President Donald Trump's (and Donald Trump, Jr.'s) misleading tweets. Facebook, for its part, has evolved from being a notorious haven

[3] *See* 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,'* THE GUARDIAN (Mar. 22, 2012), https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.
[4] *See* Arjun Kharpal, *Twitter Removes an Image Tweeted by Trump for Violating Its Copyright Policy*, CNBC (Jul. 2, 2020), https://www.cnbc.com/2020/07/02/twitter-removes-trump-image-in-tweet-for-violating-copyright-policy.html.
[5] *See* Elizabeth Dwoskin, *Trump Lashes Out at Social Media Companies After Twitter Labels Tweets with Fact Checks*, WASH. POST (May 27, 2020), https://www.washingtonpost.com/technology/2020/05/27/trump-twitter-label.

for fake news in the 2016 election cycle[6] to setting up an extensive global network of independent fact-checkers to remove and label millions of posts on its platform. This has included removing a post from President Trump's campaign account.[7] In March and April 2020, Facebook also labeled ninety million posts involving false or misleading medical information in the context of the pandemic.[8] Google has abandoned its hands-off approach to its search algorithm results and has committed to removing false political content in the context of the 2020 election[9] and to serving up prominent information by trusted health authorities in response to COVID-19 related searches on its platforms.[10]

These approaches undertaken by the major social media platforms are generally consistent with First Amendment values, both the substantive values in terms of what constitutes protected and unprotected speech, and the procedural values, in terms of process accorded to users whose speech is restricted or otherwise subject to action by the platforms. As I discuss below, the platforms have removed speech that is likely to lead to imminent harm and have generally been more aggressive in responding to medical misinformation than political misinformation. This approach tracks First Amendment substantive values, which accord lesser protection for false and misleading claims regarding medical information than for false and misleading political claims. The platforms' approaches generally adhere to First Amendment procedural values as well, including by specifying precise and narrow categories of what speech is prohibited, providing clear notice to speakers who violate their rules regarding speech, applying their rules consistently, and

[6] Olivia Solon, *2016: The Year Facebook Became the Bad Guy,* THE GUARDIAN (Dec. 12, 2016), https://www.theguardian.com/technology/2016/dec/12/facebook-2016-problems-fake-news-censorship.

[7] *See* Heather Kelly, *Facebook, Twitter Penalize Trump for Posts Containing Coronavirus Misinformation*, WASH. POST (Aug. 7, 2020, 2:25 PM), https://www.washingtonpost.com/technology/2020/08/05/trump-post-removed-facebook/.

[8] Guy Rosen, *An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19*, FACEBOOK NEWSROOM (Apr. 16, 2020), https://about.fb.com/news/2020/04/COVID-19-misinfo-update [hereinafter Rosen, *COVID-19 Update*].

[9] Zachary Evans, *Google to Place Limits on Political Advertisements Ahead of 2020 Election,* NAT'L REV. (Nov. 21, 2019), https://www.nationalreview.com/news/google-to-place-limits-on-political-advertisements-ahead-of-2020-election.

[10] Alexios Mantzarlis, *COVID-19: $6.5 Million to Help Fight Coronavirus Misinformation,* GOOGLE (Apr. 2, 2020), https://blog.google/outreach-initiatives/google-news-initiative/covid-19-65-million-help-fight-coronavirus-misinformation.

according an opportunity for affected speakers to appeal adverse decisions regarding their content.

While the major social media platforms' intervention in the online marketplace of ideas is not without its problems and not without its critics, this Article contends that this trend is by and large a salutary development—one that is welcomed by the vast majority of Americans and that has brought about measurable improvements in the online information ecosystem. Recent surveys and studies show that such efforts are welcomed by Americans[11] and are moderately effective in reducing the spread of misinformation and in improving the accuracy of beliefs of members of the public.[12] In the absence of effective regulatory measures in the United States to combat medical and political misinformation online, social media companies should be encouraged to continue to experiment with developing and deploying even more effective measures to combat such misinformation, consistent with our First Amendment substantive and procedural values.

This Article begins in Part I with a detailed examination of how each of the major social media platforms—Facebook/Instagram, Twitter, and YouTube/Google—address COVID-19 medical misinformation. Part II conducts a similar examination with respect to the platforms' evolving treatment of political misinformation, an issue made more complicated by the lack of regulation of political advertising on social media and by the practice of microtargeting of political ads. Part III assesses the platforms' measures to combat medical and political misinformation through the lens of First Amendment values, both *substantive* First Amendment values—the extent to which the platforms prioritize counterspeech over censorship, absent emergency—and *procedural* First Amendment values—the extent to which the platforms articulate (and communicate to their users) clear, neutral, and transparent rules, enforced by impartial decision-makers, with the opportunity for appeal.[13] A brief Conclusion follows.

[11] FREE EXPRESSION, HARMFUL SPEECH AND CENSORSHIP IN A DIGITAL WORLD 6 (2020), https://knightfoundation.org/wp-content/uploads/2020/06/KnightFoundation_Panel6-Techlash2_rprt_061220-v2_es-1.pdf.

[12] Lee Drutman, *Fact-Checking Misinformation Can Work. But It Might Not Be Enough,* FIVETHIRTYEIGHT (June 4, 2020), https://fivethirtyeight.com/features/why-twitters-fact-check-of-trump-might-not-be-enough-to-combat-misinformation.

[13] *See infra* Part III.

I. Platforms' Efforts to Address Medical Misinformation in the Context of the Pandemic

In recent months, arguably the most important challenge for social media platforms has been responding to the rampant spread of medical misinformation in the context of the COVID-19 pandemic. With a significant portion of the global community under some kind of lockdown order (one third of the global population by one estimate[14]), Internet connectivity—along with Internet content—are playing a more significant societal role than ever. In contrast to their previous hands-off position, the major platforms have risen to the challenge and have taken decisive action in response to medical misinformation in the context of the pandemic. The predominant focus across platforms has been on the curbing of false information, especially that which tends to encourage the spread of imminently harmful information about the virus. The platforms' actions taken in response to COVID-19-related medical misinformation have generally been more aggressive than their response to misinformation in the political arena, which is consistent with First Amendment substantive values that accord lesser protection for false and misleading statements of fact than for false and misleading political claims (as I discuss below). And the platforms' actions in the context of medical misinformation generally track First Amendment substantive values by prohibiting false and imminently harmful information. In general, the platforms have undertaken extensive measures to remove imminently harmful false medical information (e.g., posts that advocate drinking bleach to cure COVID-19), while taking less severe measures regarding less harmful or misleading medical information (e.g., posts that tout conspiracy theories claiming that Dr. Anthony Fauci created the virus), such as by labeling or reducing the reach of such posts. Although the platforms' efforts thus far are commendable, they must act much more quickly to remove harmful false and misleading medical misinformation before it goes viral, as I discuss below.

[14] Juliana Kaplan, Lauren Frias & Morgan McFall-Johnsen, *A Third of the Global Population is on Coronavirus Lockdown—Here's Our Constantly Updated List of Countries and Restrictions*, BUS. INSIDER INDIA (Jul. 11, 2020), https://www.businessinsider.in/international/news/a-third-of-the-global-population-is-on-coronavirus-lockdown-x2014-hereaposs-our-constantly-updated-list-of-countries-and-restrictions/slidelist/75208623.cms.

A. Facebook's Response to Medical Misinformation

Facebook has responded to the rampant spread of misinformation on its platform in the context of the pandemic by removing speech that it considers to be imminently harmful, while providing counterspeech in response to misleading or false speech on its platform that it deems not to be imminently harmful. Pursuant to this approach, as discussed below, Facebook has removed millions of harmful false or misleading posts related to COVID-19—including, recently, posts by President Trump, while labeling other, less harmful false or misleading posts and issuing strong warnings to those who have shared or reacted to such posts.[15] In addition, Facebook has also made prominently available its Coronavirus Information Center, a repository of curated, expert information about the virus.[16] Facebook's intent to combat medical misinformation in the context of the pandemic is commendable, but the failures in the timely implementation of its new policies are highly problematic given the degree of the public health risk. While Facebook cannot reasonably be expected to identify and curb every piece of misinformation on the pandemic on its platform, it must commit to staffing up and improving the implementation of its measures to counter medical misinformation. Facebook's success rate in curbing harmful misinformation might never be perfect, but its present approach has glaring flaws that must be remedied, as I examine below.

In a surprising move in August 2020, Facebook implemented its COVID-19 misinformation policy to delete a post from President Trump's campaign account, in which Trump can be heard saying on video, in the context of re-opening schools, that children are "almost immune" to the coronavirus.[17] While Facebook does not frequently remove medical misinformation, its Community Standards allow for removal of misinformation that contributes to the risk of physical harm or

[15] *See infra* notes 17–29 and accompanying text.

[16] Kang-Xing Jin, *Launching the Coronavirus Information Center on Facebook*, March 18, 2020, 11:06 AM update to *Keeping People Safe and Informed About the Coronavirus*, FACEBOOK NEWSROOM, https://about.fb.com/news/2020/08/coronavirus/#coronavirus-info-center (last updated Oct. 5, 2020).

[17] Kelly, *supra* note 7.

imminent violence.[18] In response to guidance from external experts, including the World Health Organization (WHO) and local health authorities, Facebook now requires the removal of "false claims about: the existence or severity of COVID-19, how to prevent COVID-19, how COVID-19 is transmitted (such as false claims that certain racial groups are immune to the virus), cures for COVID-19, and access to or the availability of essential services."[19]

Facebook has broadened its work with certified independent fact-checking organizations[20] as part of its effort to curb the spread of medical misinformation, adding eight new dedicated fact-checking partners and "expand[ing] [its] coverage to more than a dozen new countries."[21] Facebook's approach to medical misinformation generally focuses less on removing false content and more on reducing the distribution of medical misinformation once one of its independent fact-checking partners has rated it as false.[22] To the everyday user, Facebook's approach takes the form of warning displays on posts that have been deemed false, with Facebook issuing forty million such warnings in March 2020 and fifty million in April 2020.[23] Facebook claims that when people see such warning labels, "95% of the time they did not go on to view the original content."[24] In addition, in response to the pandemic, Facebook has "updated its content reviewer guidance to make clear that claims such as that people of certain races or religions have the

[18] LAURA W. MURPHY ET AL., FACEBOOK'S CIVIL RIGHTS AUDIT – FINAL REPORT 53 (2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf.

[19] *Id.*

[20] "To reduce the spread of misinformation and provide more reliable information to users, we partner with independent third-party fact-checkers globally who are certified through the non-partisan International Fact-Checking Network (IFCN)." *Partnering with Third-Party Fact-Checkers*, FACEBOOK: JOURNALISM PROJECT (Mar. 23, 2020), https://www.facebook.com/journalismproject/programs/third-party-fact-checking/selecting-partners. For a list of verified signatories, see *Verified Signatories of the IFCN Code of Principles*, https://ifcncodeofprinciples.poynter.org/signatories (last accessed Sept. 3, 2020).

[21] Rosen, *COVID-19 Update*, *supra* note 8. Facebook has also expanded the program to Instagram and now boasts "more than 60 fact-checking partners covering more than 50 languages around the world." Guy Rosen, *Investments to Fight Polarization*, FACEBOOK NEWSROOM (May 27, 2020), https://about.fb.com/news/2020/05/investments-to-fight-polarization.

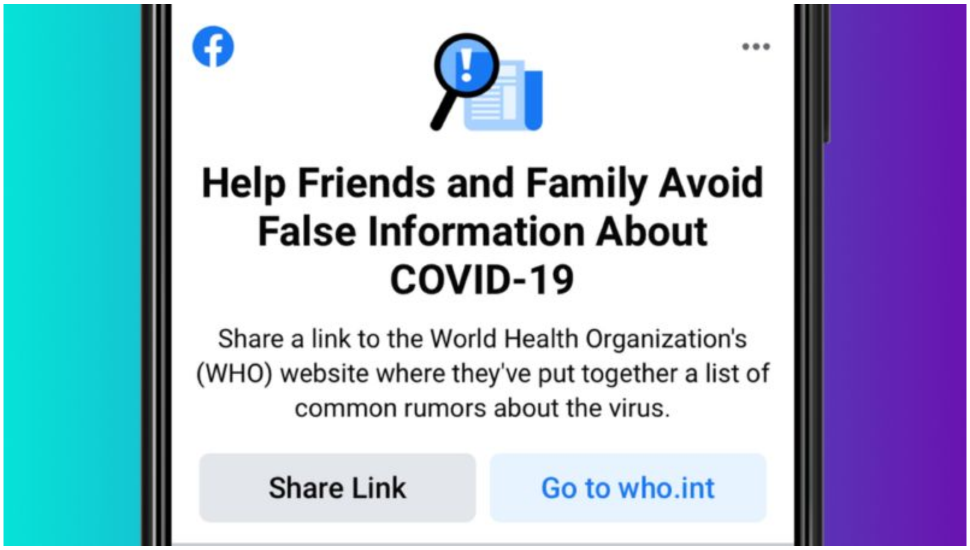
[22] Rosen, *COVID-19 Update*, *supra* note 8.

[23] *Id.*

[24] *Id.*

virus, created the virus, or are spreading the virus violate Facebook's hate speech policies."[25]

As part of its general counterspeech approach of presenting users with accurate information in response to false and misleading information, as opposed to by censoring false information, Facebook has also taken the step of reaching out to users who have interacted with (i.e., reacted to or commented on) medical misinformation related to COVID-19 and connecting those users with responses to common "myths" about COVID-19 that have been identified and addressed by the World Health Organization and inviting these users to share the link with others.[26] See notice below.



Help Friends and Family Avoid False Information About COVID-19

Share a link to the World Health Organization's (WHO) website where they've put together a list of common rumors about the virus.

Figure One: Notice on Facebook directing users who have interacted with COVID-19 misinformation to the WHO.[27]

The most common of the myths shared on Facebook tend to suggest ineffective or potentially harmful remedies for COVID-19, such as drinking bleach or disinfectant, or taking unproven and potentially harmful drugs such as

[25] MURPHY ET AL., FACEBOOK'S CIVIL RIGHTS AUDIT – FINAL REPORT, *supra* note 18 at 53.

[26] Rosen, *COVID-19 Update*, *supra* note 8.

[27] For Facebook's explanation of this notification, *see* Rosen, *supra* note 8.

hydroxychloroquine.[28] Other myths commonly seen on the platform are those claiming that measures scientifically proven to contain the spread of the virus—such as social distancing—are ineffective.[29]

In attempting to combat medical misinformation in the context of the pandemic, Facebook has also set up its Coronavirus Information Center.[30] This feature, which Facebook initially placed prominently at the top of the News Feed (so that it was immediately visible upon opening the platform) serves as a collection of relevant real-time updates about the pandemic from both national and global health authorities.[31]

Facebook's approach to combating medical misinformation on its platform[32] is heading in the right direction

[28] *Coronavirus Disease Advice for the Public: Mythbusters*, WORLD HEALTH ORG., https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters (last accessed July 21, 2020).

[29] Some recent research has highlighted the parallels in the sharing of disinformation in the current pandemic with the dissemination of information on supposed "cures" via newspapers during the 1918 flu pandemic. Suyin Haynes, *'You Must Wash Properly.' Newspaper Ads From the 1918 Flu Pandemic Show Some Things Never Change*, TIME (Mar. 27, 2020, 11:35 AM), https://time.com/5810695/spanish-flu-pandemic-coronavirus-ads/. As Elizabeth Zetland, a researcher at MyHeritage, puts it, "You were meant to cook 12 onions, get the juice and drink it the day afterwards, and that would protect you from the flu." *Id.* Newspapers were quick to urge individuals to wear or make their own masks; the Red Cross, in an ad placed in the *Daily Gazette* of Berkeley, California, called anyone not wearing a mask "a dangerous slacker." *Id.*

[30] *See* Kang-Xing Jin, *Keeping People Safe and Informed About the Coronavirus*, FACEBOOK NEWSROOM, https://about.fb.com/news/2020/08/coronavirus (last updated Oct. 5, 2020).

[31] Jin, *supra* note 16.

[32] Facebook's approach to combating medical misinformation also encompasses its response to protests involving stay-at-home measures that authorities have deemed necessary to curb the spread of the pandemic. Donnie O'Sullivan & Brian Fung, *Facebook Will Take Down Some, But Not All, Posts Promoting Anti-Stay-at-Home Protests*, CNN POLITICS (Apr. 20, 2020, 1:54 PM), https://www.cnn.com/2020/04/20/politics/facebook-COVID-shutdown-protests/index.html. Thus far, the company's response to such protests has been inconsistent. *See id.* Facebook has removed posts organizing anti-stay-at-home protests in California, New Jersey, and Nebraska after determining—in consultation with state officials—that the protests violated the states' social distancing rules. *Id.* In Pennsylvania, however, an anti-lockdown group with more than 66,000 members promoted a lockdown protest scheduled to take place in Harrisburg, without any action from Facebook. *Id.* Facebook's efforts in this area are sporadic and appear to lack a coherent strategy. In New Jersey, for example, state officials had not specifically requested that Facebook take down content promoting anti-lockdown events, but Facebook staff had been "communicating about the issue" with the governor's staff. *Id.* In Nebraska, Facebook contacted the governor's office "to learn more about Nebraska's social distancing restrictions, and the governor's staff

but is plagued by unacceptable delays. A comprehensive study undertaken by the human rights group Avaaz examined the dissemination of "over 100 pieces of misinformation . . . about the virus that were rated false and/or misleading by reputable, independent fact-checkers and that could cause public harm."[33] Avaaz's review found that "millions of the platform's users are still being put at risk," and that "the pieces of [false and/or misleading] content [sampled by Avaaz] . . . were shared over 1.7 million times on Facebook, and viewed an estimated 117 million times."[34] For example, according to Avaaz, "a harmful misinformation post that claimed that one way to rid the body of the virus is to . . . gargle with water, salt or vinegar was shared over [31,000] times before eventually being taken down after Avaaz flagged [this content for action by] Facebook."[35]

Beyond failing to apply warning labels to content that its fact-checking partners deemed to be misleading, Facebook suffers from delays in the implementation of its policies. In this age of instant digital news, unless imminently harmful medical disinformation is rapidly curbed, it runs the risk of hastening the spread of the virus.[36] And yet according to Avaaz, "it can take up to 22 days for the platform to downgrade [false and/or misleading content related to the virus] and issue warning labels."[37] The lag is even more severe in the case of non-English content, where "[o]ver half (51%) of non-English misinformation content had no warning labels."[38] Fortunately, Facebook seems

provided publicly available information about Nebraska's 10-person limit and directed health measures." *Id.* Facebook is apparently reaching out to governments on these matters because "[u]nless government prohibits the event during this time, we allow it to be organized on Facebook." *Id.* Yet, at least in the case of Nebraska, Facebook's effort seems to be a somewhat fumbling one, with Facebook employees reaching out to state officials to learn about information that is already readily available. *See id.* Facebook's hesitance to remove posts related to protests likely stems from a deeper worry about becoming the online policeman on the question of the constitutional right to assemble.

[33] *How Facebook Can Flatten the Curve of the Coronavirus Infodemic*, AVAAZ at 2 (Apr. 15, 2020), https://avaazimages.avaaz.org/facebook_coronavirus_misinformation.pdf [hereinafter *How Facebook Can Flatten the Curve*].

[34] *Id.*

[35] *Id.* at 10 ("2,611 clones [of the same false post] remain on the platform with over 92,246 interactions. Most of these cloned posts have no warning labels from Facebook.").

[36] *See* Robyn Caplan, *COVID-19 Misinformation Is a Crisis of Content Mediation*, BROOKINGS: TECHSTREAM (May 7, 2020), https://www.brookings.edu/techstream/covid-19-misinformation-is-a-crisis-of-content-mediation.

[37] *How Facebook Can Flatten the Curve*, *supra* note 33, at 2.

[38] *Id.* at 3.

willing to change course, as evidenced by its willingness to institute a retroactive alert system, whereby each user exposed to harmful misinformation will be notified and provided with accurate information.[39] Avaaz indicated that members of Facebook's misinformation team made such a commitment in a conversation with Avaaz staff in April 2020.[40]

B. Twitter's Response to Medical Misinformation

Twitter's response to medical misinformation in the context of the pandemic—like its response to false political ads, as discussed below—has been more forceful than Facebook's approach. Twitter's response includes removing harmful posts containing medical misinformation that "could directly pose a risk to people's health or well-being," counterspeech through labelling other posts containing medical misinformation, and also directing users to truthful and accurate information about the pandemic.[41] Twitter has also limited the functionality of user accounts that violate its policies.[42]

Under a recent update to Twitter's rules, tweets that cause harm, including in the context of the pandemic, will be removed.[43] Accordingly, Twitter will remove from its platform tweets along the lines of "social distancing is not effective" or "the news about washing your hands is propaganda."[44] An important component of Twitter's effort includes broadening its definition of harmful tweets, so as to more proactively target and remove content that expressly contradicts the most up-to-date guidance from authoritative health sources such as the Center for Disease Control and the World Health Organization.[45]

[39] *Id.* at 2.
[40] *See id.* at 19 n. 5.
[41] *Coronavirus: Staying Safe and Informed on Twitter*, TWITTER BLOG (Apr. 3, 2020), https://blog.twitter.com/en_us/topics/company/2020/covid-19.html.
[42] *See* Twitter Comms (@TwitterComms), TWITTER (July 28, 2020, 10:15 AM), https://twitter.com/TwitterComms/status/1288115957005578246.
[43] *Coronavirus: Staying Safe and Informed on Twitter*, *supra* note 41.
[44] Jack Morse, *Twitter Steps up Enforcement in the Face of Coronavirus Misinformation*, MASHABLE (Mar. 18, 2020), https://mashable.com/article/twitter-cracks-down-coronavirus-misinformation/.
[45] Vijaya Gadde & Matt Derella, *An Update on Our Continuity Strategy During COVID-19*, TWITTER BLOG (Mar. 16, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html. For a full list of Twitter's pandemic-related misinformation policies, see *id.*

Twitter states that it is working with "trusted partners, including public health authorities . . . and governments" to both identify and remove harmful medical misinformation.[46] Twitter states it will remove a broad range of content. Twitter is focused on tweets that contain a "denial of global or local health authority recommendations . . . with the intent to influence people into acting against recommended guidance," those that "[describe] alleged cures for COVID-19, which are not immediately harmful but are known to be ineffective . . . ," those that describe "harmful treatments or protection measures which are known to be ineffective . . . ," and those that deny "established scientific facts about transmission during the incubation period."[47]

Twitter is also targeting tweets that go beyond medical misinformation and appear to encourage societal unrest. For example, the company states that it will target and remove tweets that contain "[s]pecific and unverified claims that incite people to action and cause widespread panic, social unrest or large-scale social disorder," as well as tweets that contain "[s]pecific and unverified claims made by people impersonating a government or health official or organization."[48] One such example was a "parody account of an Italian health official stating that the country's quarantine was over."[49]

In May 2020, Twitter announced a policy of placing warning labels on tweets containing misinformation related to COVID-19, including tweets that are issued by world leaders.[50] According to Twitter's head of site integrity, Yoel Roth, such warning labels will apply to "anyone sharing misleading information that meets the requirements of Twitter's policy, and no exceptions will be made for the tweets of world leaders."[51]

[46] *Id.*
[47] *Id.*
[48] *Id.*
[49] *Id.*
[50] Yoel Roth & Nick Pickles, *Updating Our Approach to Misleading Information*, TWITTER BLOG (May 11, 2020), https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.
[51] Yoel Roth (@yoyoel), TWITTER (May 11, 2020, 3:10 PM), https://twitter.com/yoyoel/status/1259923758522855426.

Pursuant to its policy of removing COVID-19 related medical misinformation, including from world leaders, in August 2020 Twitter required the Trump campaign to remove a tweet in which Trump claimed that children are "almost immune" to the virus.[52] Twitter suspended the account's tweeting privileges until the post was deleted, citing its rules on COVID-19.[53]

In addition to removing and labeling tweets in its attempts to restrict false and misleading medical information in the context of the pandemic, Twitter is also pursuing other means, including restricting the functionality of Twitter accounts that spread such information. For example, on July 28, Twitter penalized Donald Trump, Jr., for posting misinformation in the form of a Breitbart video showing a group of doctors making misleading and false claims about the COVID-19 pandemic.[54] In the video, a group of people dressed in white lab coats, who call themselves "America's Frontline Doctors," staged a press conference in front of the U.S. Supreme Court and claimed that hydroxychloroquine is "a cure for Covid" and people "don't need a mask to slow the spread of coronavirus."[55] Twitter ordered Trump Jr. to delete this misleading tweet, added a note to its trending topics warning about the potential dangers of hydroxychloroquine, and took measures to limit Trump Jr.'s account functionality for twelve hours.[56] Facebook and YouTube also removed the offending video, but not before it had been viewed millions of times.[57]

[52] Shannon Bond, *Twitter, Facebook Remove Trump Post Over False Claim About Children And COVID-19*, NPR (Aug. 5, 2020, 8:49 PM), https://www.npr.org/2020/08/05/899558311/facebook-removes-trump-post-over-false-claim-about-children-and-covid-19.

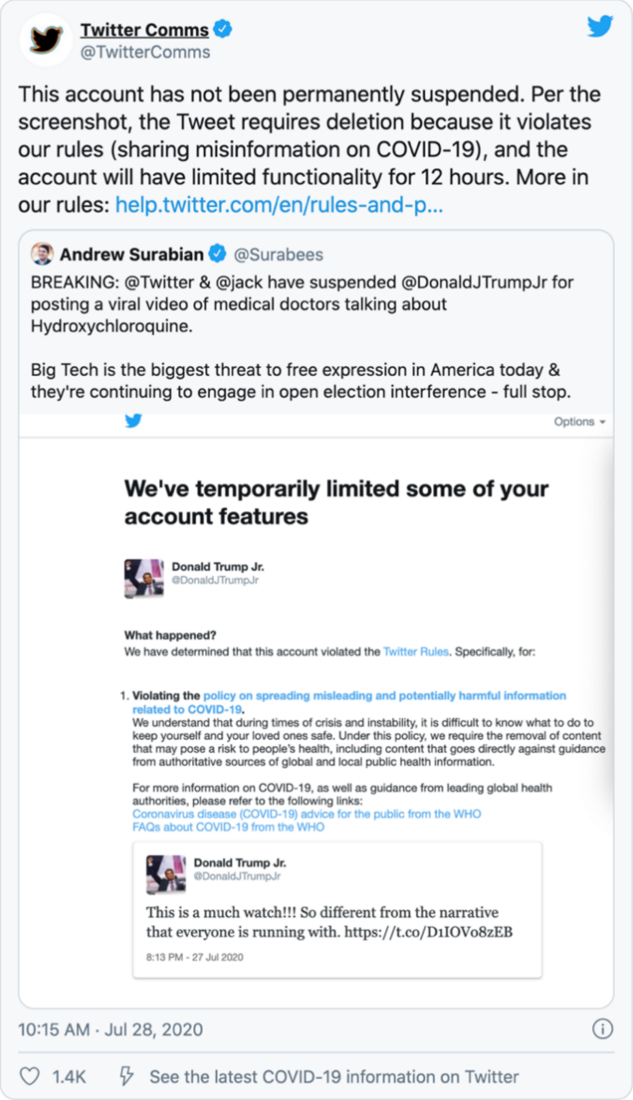
[53] Roth, *supra* note 51.

[54] *See* Rachel Lerman, Katie Shepherd & Taylor Telford, *Twitter Penalizes Donald Trump Jr. for Posting Hydroxychloroquine Misinformation Amid Coronavirus Pandemic*, WASH. POST (July 28, 2020), https://www.washingtonpost.com/nation/2020/07/28/trump-coronavirus-misinformation-twitter/.

[55] *See* Sam Shead, *Facebook, Twitter and YouTube Pull 'False' Coronavirus Video After It Goes Viral,* CNBC (July 28, 2020, 7:37 AM), https://www.cnbc.com/2020/07/28/facebook-twitter-youtube-pull-false-coronavirus-video-after-it-goes-viral.html.

[56] *See* Lerman, Shepherd & Telford, *supra* note 54.

[57] *See* Shead, *supra* note 55.



Twitter Comms
@TwitterComms

This account has not been permanently suspended. Per the screenshot, the Tweet requires deletion because it violates our rules (sharing misinformation on COVID-19), and the account will have limited functionality for 12 hours. More in our rules: help.twitter.com/en/rules-and-p...

Andrew Surabian @Surabees

BREAKING: @Twitter & @jack have suspended @DonaldJTrumpJr for posting a viral video of medical doctors talking about Hydroxychloroquine.

Big Tech is the biggest threat to free expression in America today & they're continuing to engage in open election interference - full stop.

We've temporarily limited some of your account features

Donald Trump Jr.
@DonaldJTrumpJr

What happened?
We have determined that this account violated the Twitter Rules. Specifically, for:

1. Violating the policy on spreading misleading and potentially harmful information related to COVID-19.
We understand that during times of crisis and instability, it is difficult to know what to do to keep yourself and your loved ones safe. Under this policy, we require the removal of content that may pose a risk to people's health, including content that goes directly against guidance from authoritative sources of global and local public health information.

For more information on COVID-19, as well as guidance from leading global health authorities, please refer to the following links:
Coronavirus disease (COVID-19) advice for the public from the WHO
FAQs about COVID-19 from the WHO

Donald Trump Jr.
@DonaldJTrumpJr

This is a much watch!!! So different from the narrative that everyone is running with. https://t.co/D1IOVo8zEB

8:13 PM - 27 Jul 2020

10:15 AM · Jul 28, 2020

Figure Two: Twitter's explanation of limiting the functionality of Donald Trump Jr.'s account in July 2020.[58]

In addition to implementing systems to remove false and misleading medical information, Twitter is also prominently featuring truthful and accurate information about COVID-19 through its "Know The [F]acts" search prompt.[59] In early 2020, Twitter expanded its #KnowTheFacts program, which it had earlier put in place to help the public find credible information

[58] Twitter Comms (@TwitterComms), TWITTER (July 28, 2020, 10:15 AM), https://twitter.com/TwitterComms/status/1288115957005578246.
[59] *Coronavirus: Staying Safe and Informed on Twitter*, *supra* note 41.

on immunization and vaccine health.[60] The purpose of this program was to surface and "highlight credible information" on the virus.[61] The program also ensured that when Twitter users access the platform to search for information about the virus, they are first met with "credible, authoritative information" from reliable sources.[62] A further component of Twitter's #KnowTheFacts program limits auto-suggest results that may direct Twitter users to misinformation on Twitter.[63] And, similar to Facebook, Twitter has created a specific webpage dedicated to providing the latest authoritative information on the pandemic.[64] This resource—Twitter's "COVID-19 Event Page"—provides an aggregation of credible news updates on the pandemic, curated with content from verified sources like *The New York Times*, *Associated Press*, and *Reuters*, as well as public health sources such as the Center for Disease Control (CDC), which provide relevant virus-related updates.[65]

Twitter's aggressive approach to medical misinformation on its platform appears to be extensive and effective, so far. According to Twitter's reporting on the implementation of its policies regarding medical misinformation in the context of the pandemic, it has removed thousands of tweets containing misleading and potentially harmful content and has "challenged more than 1.5 million accounts which were targeting discussions around COVID-19 with spammy or manipulative behaviors."[66]

C. Google/YouTube's Response to Medical Misinformation

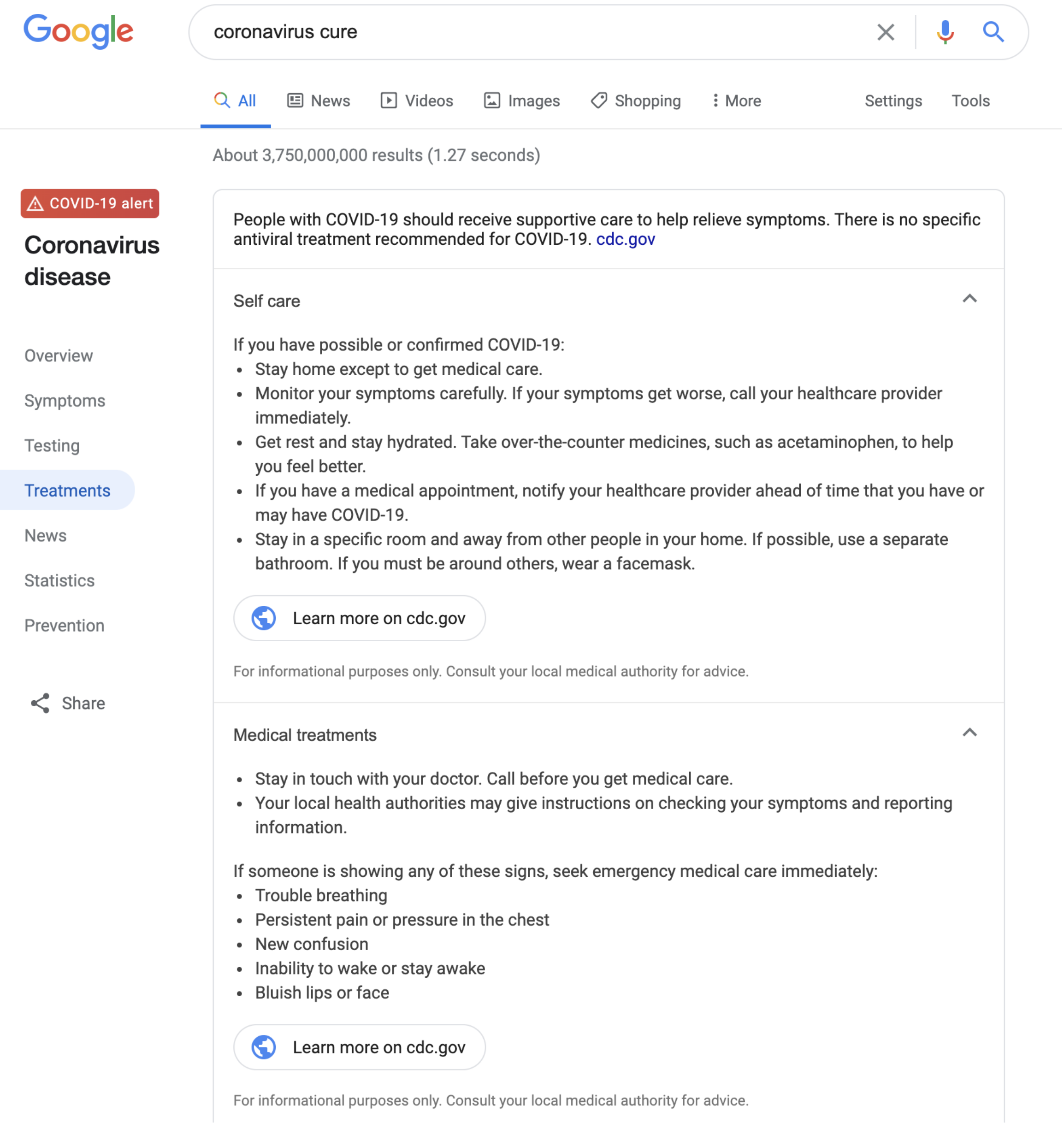
Google's approach to searches that seek information on COVID-19 is similarly proactive. Consistent with First Amendment values, Google/YouTube has avoided censorship of medical misinformation by employing counterspeech in the form of directing users to authoritative sources when they search for terms likely to produce search results containing

[60] *Id.*
[61] *Id.*
[62] *Id.*
[63] *Id.*
[64] *See COVID-19: Latest News Updates from Around the World*, TWITTER, https://twitter.com/i/events/1219057585707315201 (last visited Sept. 3, 2020). Most tweets and updates are from verified news accounts such as *The New York Times*, *Associated Press*, and *Reuters*, as well as public health sources such as the CDC and similar international entities. *See id.*
[65] *Id.*
[66] *Coronavirus: Staying Safe and Informed on Twitter*, *supra* note 41.

misinformation. Google/YouTube also removed one major incentive to post such content in the first place by demonetizing videos that violate its policies, in lieu of censoring the content outright.

Google has, for most of its history, deferred solely to its complex algorithms to produce search results without human intervention. However, the company now has taken the approach of having searches related to coronavirus trigger a type of "SOS alert," resulting in prominent displays of news from "mainstream publications including National Public Radio, followed by information from the U.S. CDC and the WHO."[67]

[67] Mark Bergen & Garrit De Vynck, *Google Scrubs Coronavirus Misinformation on Search, YouTube*, BLOOMBERG (Mar. 10, 2020, 6:00 AM), https://www.bloomberg.com/news/articles/2020-03-10/dr-google-scrubs-coronavirus-misinformation-on-search-youtube.



coronavirus cure

About 3,750,000,000 results (1.27 seconds)

Coronavirus disease

People with COVID-19 should receive supportive care to help relieve symptoms. There is no specific antiviral treatment recommended for COVID-19. cdc.gov

Self care

If you have possible or confirmed COVID-19:

- Stay home except to get medical care.
- Monitor your symptoms carefully. If your symptoms get worse, call your healthcare provider immediately.
- Get rest and stay hydrated. Take over-the-counter medicines, such as acetaminophen, to help you feel better.
- If you have a medical appointment, notify your healthcare provider ahead of time that you have or may have COVID-19.
- Stay in a specific room and away from other people in your home. If possible, use a separate bathroom. If you must be around others, wear a facemask.

For informational purposes only. Consult your local medical authority for advice.

Medical treatments

- Stay in touch with your doctor. Call before you get medical care.
- Your local health authorities may give instructions on checking your symptoms and reporting information.

If someone is showing any of these signs, seek emergency medical care immediately:

- Trouble breathing
- Persistent pain or pressure in the chest
- New confusion
- Inability to wake or stay awake
- Bluish lips or face

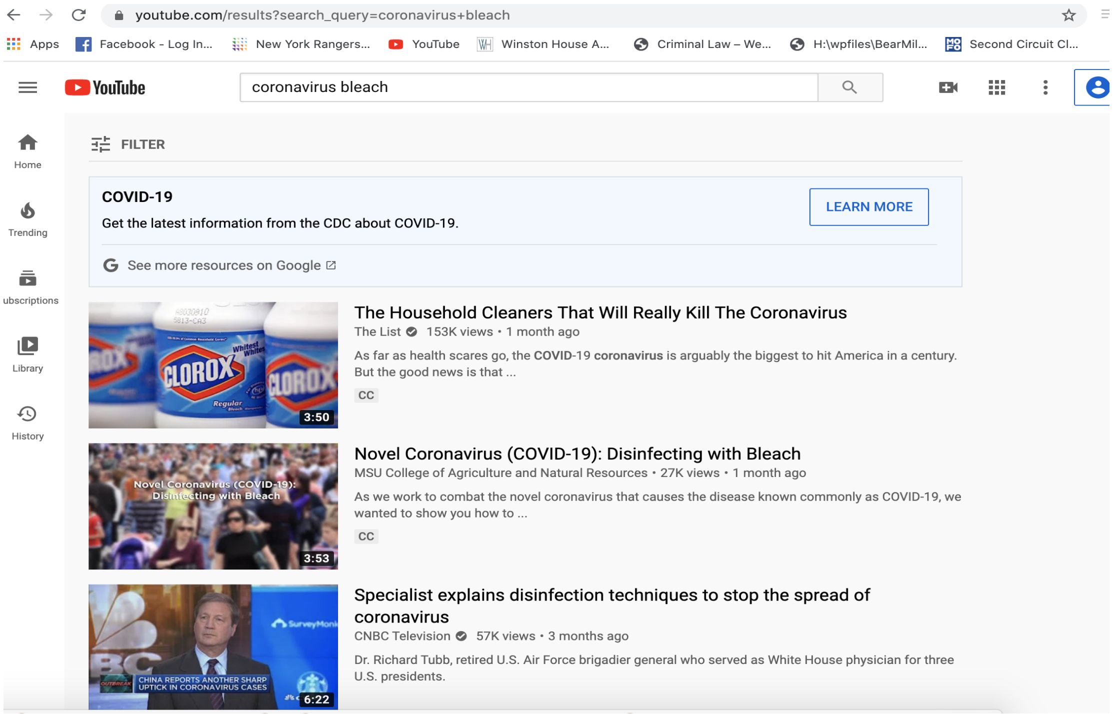
For informational purposes only. Consult your local medical authority for advice.

Figure Three: Google search results page in response to searching for "coronavirus cure," as of July 6, 2020.[68]

In addition, YouTube has modified its Terms of Service to prohibit any content that directly contradicts advice from the

[68] See, e.g., *Coronavirus Cure*, GOOGLE, https://www.google.com/search?client=firefox-b-1-d&q=coronavirus+cure (last accessed Jan. 12, 2021) (displaying an informational banner linking to the CDC in response to the search "coronavirus cure").

WHO.[69] In an update to its monetization policy, YouTube announced that it will prohibit videos that seek to capitalize on coronavirus-related conspiracies.[70] Instead, it has directed users to videos debunking the conspiracies.[71]



COVID-19

Get the latest information from the CDC about COVID-19.

See more resources on Google

The Household Cleaners That Will Really Kill The Coronavirus
The List 153K views • 1 month ago
As far as health scares go, the COVID-19 **coronavirus** is arguably the biggest to hit America in a century. But the good news is that ...

Novel Coronavirus (COVID-19): Disinfecting with Bleach
MSU College of Agriculture and Natural Resources • 27K views • 1 month ago
As we work to combat the novel coronavirus that causes the disease known commonly as COVID-19, we wanted to show you how to ...

Specialist explains disinfection techniques to stop the spread of coronavirus
CNBC Television 57K views • 3 months ago
Dr. Richard Tubb, retired U.S. Air Force brigadier general who served as White House physician for three U.S. presidents.

Figure Four: YouTube search results page in response to searching for "coronavirus bleach," as of July 6, 2020.[72]

YouTube, however, like Facebook, has run into difficulties countering medical misinformation on its platform, especially in regions where fact-checking is not as readily achievable or practical as it is in the United States.[73] This is partly

[69] *Coronavirus: YouTube Bans 'Medically Unsubstantiated' Content*, BBC NEWS (Apr. 22, 2020), https://www.bbc.com/news/technology-52388586.

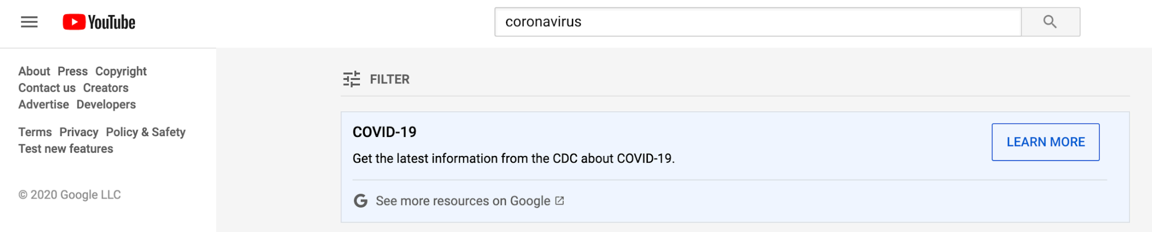
[70] *Monetization update on COVID-19 content,* YOUTUBE HELP, https://support.google.com/youtube/answer/9803260?hl=en (last accessed Oct. 16, 2020).

[71] *Health information panels,* YOUTUBE HELP, https://support.google.com/youtube/answer/9795167 (last accessed Oct. 16, 2020).

[72] *See, e.g.*, *Coronavirus Bleach*, YOUTUBE, https://www.youtube.com/results?search_query=coronavirus+bleach (last visited Jan. 12, 2021) (displaying a warning banner and linking to the CDC in response to a search for "coronavirus bleach").

[73] Ryan Browne, *YouTube Expands Fact-checking Feature for Video Searches to Europe*, CNBC (Sept. 24, 2020), https://www.cnbc.com/2020/09/24/youtube-expands-fact-checking-feature-for-video-searches-to-europe.html.

a result of the nature of the medium itself. YouTube videos typically involve creative elements such that the question of whether the content is true or false becomes more complex. YouTube has recently adopted the same approach as Google, providing a banner at the top of searches for terms such as "coronavirus" or "coronavirus cure" with a link to the CDC's official page (see below).[74]

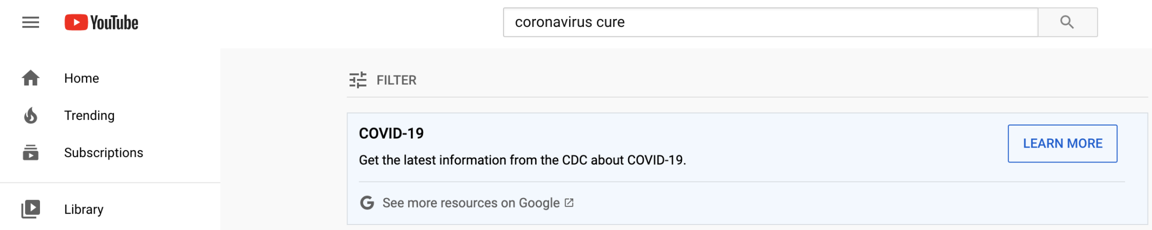


coronavirus

COVID-19

Get the latest information from the CDC about COVID-19.

Figure Five: YouTube search results page in response to searching for "coronavirus," as of July 6, 2020.[75]



coronavirus cure

COVID-19

Get the latest information from the CDC about COVID-19.

Figure Six: YouTube search results page in response to searching for "coronavirus cure," as of July 6, 2020.[76]

[74] *See*, *e.g.*, *Coronavirus*, YOUTUBE, https://www.youtube.com/results?search_query=Coronavirus (last visited Jan. 12, 2021) (displaying a warning banner and linking to the CDC in response to a search for "coronavirus").

[75] *Id.*

[76] *See*, *e.g.*, *Coronavirus Cure*, YOUTUBE, https://www.youtube.com/results?search_query=coronavirus+cure (last visited Jan. 12, 2021) (displaying a warning banner and linking to the CDC in response to a search for "coronavirus cure").

II. PLATFORMS' EFFORTS TO ADDRESS POLITICAL MISINFORMATION: MEASURES TO COMBAT FALSE AND MISLEADING POLITICAL SPEECH AND MICROTARGETING OF POLITICAL ADS

A. Introduction

While the severe consequences of medical misinformation in the pandemic context are patently clear—more people will be more likely to contract the disease and/or suffer related medical harms—the consequences of political misinformation were evidenced in the aftermath of the United States' 2016 presidential election—namely, the targeted suppression of certain demographics of voters[77] and rampant disinformation injected into the public discourse by foreign operatives.[78] The presence of political misinformation on social media introduces two new complications: the host of regulations that apply to traditional broadcasting of political advertisements do not apply to social media platforms,[79] and the microtargeting[80] of political advertisements threatens the broad exposure and public scrutiny that are necessary for the marketplace of ideas[81] to function.

Today's online information ecosystem continues to be a forum for political and election-related misinformation, as it was four years ago in the context of the 2016 election. Misinformation and disinformation on the Internet are particularly problematic given that the Internet is a dominant (if not *the* dominant) source of information in the political sphere, with two-thirds of Americans identifying Internet sources as their leading sources of information in connection with the 2016 U.S. presidential election.[82] In addition, misinformation can spread

[77] Spencer Overton, *State Power to Regulate Social Media Companies to Prevent Voter Suppression*, 53 U.C. DAVIS L. REV. 1793, 1795–98 (2020).

[78] *Hearing on Social Media Influence in the 2016 United States Elections, Before the Senate Select Comm. on Intel.,* U.S. SENATE SELECT COMM. INTEL. (2017) https://www.intelligence.senate.gov/hearings/open-hearing-social-media-influence-2016-us-elections#.

[79] *See infra* Part II.C.

[80] *See infra* Part II.E.

[81] *See generally* Dawn C. Nunziato, *The Marketplace of Ideas Online*, 94 NOTRE DAME L. REV. 1519 (2019) [hereinafter Nunziato, *The Marketplace of Ideas Online*].

[82] *See* Honest Ads Act, S. 1989, 115th Cong. § 3(10) (2017); ELECTION 2016: CAMPAIGNS AS A DIRECT SOURCE OF NEWS 28 (2016) https://www.journalism.org/wp-content/uploads/sites/8/2016/07/PJ_2016.07.18_election-2016_FINAL.pdf.

faster and farther than truthful information on social media.[83] 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 1% 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.[84] The authors of the study in *Science* 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 3 million people more than 4.5 million times."[85] 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."[86]

False and misleading political content on social media platforms—especially on Facebook—played a significant role in influencing members of the electorate leading up to the 2016 election. "More than one quarter of voting-age adults visited a false news website . . . in the final weeks of the 2016 campaign."[87]

[83] *See* Soroush Vosoughi, Deb Roy, & Sinan Aral, *The Spread of True and False News Online*, 359 SCIENCE 1146, 1146–1151 (2018).
[84] *Id.* at 1148.
[85] The authors of the study in Science classified news as "true" or "false" using information from six independent fact-checking organizations that exhibited ninety-five percent to ninety-eight percent agreement on the classifications. *Id.* at 1146.
[86] *Id.* Interestingly, the authors further observe that "[c]ontrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it." *Id.* But that is not to diminish the role that bots played in Russian interference in the 2016 election. Foreign interference in our 2016 presidential elections was clearly exacerbated by the use of automation in the form of bots, trolls, and fake accounts and by the use of microtargeted political advertisements to amplify disinformation, manipulate public discourse, exacerbate political and social divisions, and deceive voters on a mass scale, especially via Twitter's platform, in a manner that was targeted to members of the U.S. electorate, especially in swing states. Natasha Bertrand, *Twitter Users Spreading Fake News Targeted Swing States in the Run-Up to Election Day*, BUS. INSIDER (Sept. 28, 2017, 1:17 PM), https://www.businessinsider.com/fake-news-and-propaganda-targeted-swing-states-before-election-2017-9. Trump won the Electoral College because some eighty thousand votes went his way in Wisconsin, Michigan, and Pennsylvania. *See e.g.*, KATHLEEN HALL JAMIESON, CYBERWAR: HOW RUSSIAN HACKERS AND TROLLS HELPED ELECT A PRESIDENT 67 (2018).
[87] Danielle Kurtzleben, *Did Fake News on Facebook Help Elect Trump? Here's What We Know*, NPR (Apr. 11, 2018), https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election.

Indeed, "[i]n the months leading up to the election, the top 20 fake news stories had more "engagements" (which includes shares, reactions, and comments) than the top twenty hard news stories—approximately nine million engagements with fake news as compared to about seven million engagements with hard news stories.[88] According to Buzzfeed, "[i]n the final three months of the U.S. presidential campaign, the top-performing fake election stories on Facebook generated more engagement than the top news stories from major news outlets like *The New York Times*, *The Washington Post*, *Huffington Post*, and *NBC News*."[89]

B. Political Speech and Political Advertising on Social Media Platforms Today

Political advertising on social media platforms is big business—and, as of this writing—still largely *unregulated* business in the United States. The total amount spent on digital political advertising in the U.S. is expected to reach $2.9 billion in 2020 (an increase of over 100% from 2016), with Google and Facebook capturing the vast majority of digital political advertising.[90] Because of the power of such ads in influencing our democratic processes, the use of microtargeting to target specific, narrow segments of the electorate, and the substance of such ads—including false and misleading information—have been subject to intense scrutiny, as discussed below.

Because political advertising has the potential to affect our democratic processes in powerful ways, it has traditionally been subject to a host of government regulations, including transparency regulations, disclosure regulations, public file regulations, and prohibitions on foreign participation.[91] Yet, as of this writing, such government regulations only apply to traditional media.[92] Despite the fact that digital advertising has

[88] *Id.*

[89] Craig Silverman, *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook*, BUZZFEED NEWS (Nov. 16, 2016, 5:15 PM), https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook#.emA15rzd0.

[90] Emily Glazer, *Facebook Weighs Steps to Curb Narrowly Targeted Political Ads*, WALL STREET J. (Nov. 21, 2019), https://www.wsj.com/articles/facebook-discussing-potential-changes-to-political-ad-policy-11574352887.

[91] *See infra* Part II.C.

[92] *See infra* Parts II.C–D.

surpassed advertising in print and television, government regulations on political advertising generally do not apply to online mediums.[93] Although the federal Honest Ads Act was introduced as an attempt to remedy this regulatory gap and to extend this host of regulations on political advertising to social media platforms, as of this writing, the Act has not been enacted into law.[94] In addition, Maryland's attempts to regulate political advertising online have been subject to successful First Amendment challenges, as discussed below.[95] Accordingly, the social media giants—Facebook, Twitter, and Google—have been left to their own devices to determine whether and how to regulate political advertising and the microtargeting of political ads on their platforms.

Below, in Part C, I briefly survey the current state of the regulation of political advertising applicable to traditional mediums of expression. In Part D, I examine the proposed federal Honest Ads Act. In Part E, I turn to the special problems of microtargeting of political advertising on social media. I then analyze the steps that the major social media platforms have taken—and have declined to take—to address the problems caused by false political speech on their forums and by the microtargeting of political ads in particular.

C. Federal Regulation of Political Advertising Applicable to Traditional Media

Various federal statutes, Federal Election Commission (FEC) rules, and Federal Communications Commission (FCC) rules currently impose transparency requirements on political advertisements disseminated by broadcast, cable, and satellite providers, and also prohibit these providers from accepting foreign advertisements in U.S. elections. First, FEC regulations

[93] Kurt Wagner, *Digital Advertising in the US is Finally Bigger Than Print and Television*, VOX MEDIA (Feb. 20, 2019, 9:02 AM), https://www.vox.com/2019/2/20/18232433/digital-advertising-facebook-google-growth-tv-print-emarketer-2019.

[94] *Actions Overview S. 1989 – 115th Congress (2017-2018)*, CONGRESS.GOV, https://www.congress.gov/bill/115th-congress/senate-bill/1989/actions (last accessed Jan. 12, 2021).

[95] In December 2019, the U.S. Court of Appeals for the Fourth Circuit held that two of Maryland's regulations for political advertising online violated the First Amendment as applied to a group of media plaintiffs, including *The Washington Post* and *The Baltimore Sun*, among others. Wash. Post v. McManus, 944 F.3d 506, 520 (4th Cir. 2019).

impose transparency requirements on political advertisements disseminated via non-social media: "any public communication made by a political committee—including communications that do not expressly advocate the election or defeat of a clearly identified federal candidate or solicit a contribution—must display a disclaimer."[96] Additionally, "[d]isclaimers must also appear on political committees' internet websites and in certain email communications."[97] All electioneering communications, public communications that expressly advocate the election or defeat of a clearly identified candidate, and public communications that solicit a contribution require a disclaimer, regardless of who has paid for them.[98] Furthermore "[p]ublic communications include electioneering communications and any other form of general public political advertisement, including communications made using the following media: broadcast, cable or satellite; newspaper or magazine; outdoor advertising facility; mass mailing (more than 500 substantially similar mailings within 30 days); phone bank (more than 500 substantially similar calls within 30 days); [and] communications placed for a fee on another person's website."[99]

Second, the "Foreign Participation Ban" prohibits foreign nationals from attempting to influence elections through donations, expenditures, or other things of value.[100] Existing regulations applicable to broadcast, cable, and satellite platforms include a broad prohibition on the involvement of foreign nationals with elections in the United States.[101] Foreign nationals are prohibited from making any contribution, donation, or expenditure in connection with any federal, state, or local election; making any contribution or donation to any committee or organization of any national, state, or local political party; or making any disbursement for an electioneering communication.[102]

[96] *Advertising and Disclaimers*, FED. ELECTION COMM'N, https://www.fec.gov/help-candidates-and-committees/making-disbursements/advertising/ (last accessed Jul. 19, 2020). For a definition of "[p]ublic [c]ommunication," see 52 U.S.C. § 30101(22).
[97] *Advertising and Disclaimers*, *supra* note 96.
[98] *Id.*
[99] *Id.*
[100] 52 U.S.C. § 30121(a)(1)(A), (C).
[101] *See id.* § 30121(a).
[102] *See id.* § 30121(a)(1).

Third, the Bipartisan Campaign Reform Act (BCRA), which applies to traditional media, imposes disclosure and public file requirements aimed at informing the electorate about the source of election related advertisements; these provisions have been upheld by the Supreme Court.[103] BCRA § 311 requires that televised "electioneering communications" funded by anyone other than a candidate include a statement clearly indicating who was responsible for the ad, along with the name and address (or web address) of the person who funded the ad.[104] In addition, BCRA requires that anyone who spent more than $10,000 on electioneering communications within a calendar year file a detailed statement with the FEC, providing their name, the amount of the expenditure, and the name of the election to which the communication was directed, among other details.[105] In upholding BCRA's disclosure and public file requirements against a First Amendment challenge by Citizens United, the Supreme Court explained that these provisions "provid[e] the electorate with information" and "insure that voters are fully informed about the person or group who is speaking . . . so that people will be able to evaluate the arguments to which they are being subjected."[106] The Court concluded that these requirements were a less restrictive alternative compared to other, more extensive regulations of political speech, since "the public has an interest in knowing who is speaking about a candidate shortly before an election," and that this "informational interest alone is sufficient to justify application of [the Act] to these ads."[107]

None of the above regulations currently apply to political advertising on social media.

D. The Proposed Honest Ads Act

As discussed above, various federal statutes and Federal Election Commission rules currently impose transparency requirements on political advertisements disseminated by

[103] *See* Bipartisan Campaign Reform Act of 2002 (BCRA), Pub. L. No. 107–155, 116 Stat. 81 (2002); *see also* Citizens United v. Fed. Election Comm'n, 558 U.S. 310 (2010); Buckley v. Valeo, 424 U.S. 1 (1976).

[104] BCRA § 311.

[105] BCRA § 201.

[106] *Citizens United*, 558 U.S. at 368 (internal citations omitted).

[107] *Id.* at 369.

broadcast, cable, and satellite providers, and also impose requirements on these providers prohibiting foreign participation in U.S. elections.[108] Social media platforms like Google, Facebook, and Twitter are currently not subject to the federal statutes and FEC rules[109] discussed above.

The Honest Ads Act, introduced in October 2017 by Senators Mark Warner (D–Virginia), Amy Klobuchar (D–Minnesota), and the late John McCain (R–Arizona), seeks to remedy this regulatory disparity.[110] It imposes transparency regulations on online political advertisements and requires that online platforms enforce the longstanding ban on foreign participation in United States elections.[111] Although, as discussed *infra*, social media platforms like Twitter, Google, and Facebook are undertaking substantial measures themselves to address such problems,[112] these measures may be revisited or revoked by the platforms at any time. Therefore, government regulation in the form of the Honest Ads Act is still an important tool for addressing these problems, and, indeed, one that is welcomed by the platforms.[113]

E. The Special Problems Caused by Microtargeting of Political Ads

Microtargeting of ads on social media platforms is a practice that generally allows advertisers to limit their messaging to narrow 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.[114] Microtargeting of ads in general stands in sharp contrast to the broadcasting of ads in mediums like major metropolitan newspapers, radio and television, through which advertisers provide content to a broad audience (e.g., to all readers of *The*

[108] *See supra* Part II.C.

[109] *See generally* Abby K. Wood & Ann M. Ravel, *Fool Me Once: Regulating "Fake News" and Other Online Advertising*, 91 S. CAL. L. REV. 1223 (2018).

[110] *See* S.1989, 115th Cong. (2017).

[111] *See* S.1989 §§ 3–4.

[112] *See infra* Part II.F.

[113] Both Facebook and Twitter have come out in support of the Honest Ads Act. *See* Aimee Picchi, *Facebook: What Is the Honest Ads Act?*, CBS NEWS (Apr. 11, 2018, 1:25 PM), https://www.cbsnews.com/news/facebook-hearings-what-is-the-honest-ads-act; Twitter Public Policy (@Policy), TWITTER (Apr. 10, 2018, 11:54 AM), https://twitter.com/Policy/status/983734917015199744.

[114] *Microtargeting*, INFO. COMM'R OFF. (Oct. 16, 2020), https://ico.org.uk/your-data-matters/be-data-aware/social-media-privacy-settings/microtargeting/.

Washington Post). Microtargeting 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).[115] This practice is essentially the "online equivalent of whispering millions of different messages into zillions of different ears for maximum effect and with minimum scrutiny."[116] It 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.[117] Facebook, for example, reportedly tracks a list of over 1,100 attributes for each user, including information regarding users' demographics, behaviors, and interests.[118]

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."[119] While businesses derive certain benefits from the microtargeting of ads in nonpolitical contexts, microtargeting of

[115] *See, e.g.*, 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/.

[116] 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.

[117] *Id.*

[118] *See* Till Speicher et al., *Potential for Discrimination in Online Targeted Advertising*, 81 PROC. MACHINE LEARNING RES. 1, 7 (2018) ("For each user in the U.S., 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 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)).

[119] Ellen L. Weintraub, *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/.

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.[120] 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.

The microtargeting of political ads, compared to the dissemination of political ads via traditional media outlets, is problematic for a number of reasons from a free speech perspective. First, political ads disseminated via traditional media are subject to a host of federal regulations requiring transparency, disclosure, limitations on foreign interference, etc., as discussed above, whereas ads disseminated via social media are not.[121] Second, 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, are not similarly subject to broad exposure or broad public scrutiny. Third, 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' content . . . further encourages the success of misinformation campaigns."[122]

In short, the microtargeting of political ads disseminated via social media is especially pernicious because it is not subject to regulatory scrutiny, not subject to meaningful widespread

[120] *See generally* Nunziato, *The Marketplace of Ideas Online*, *supra* note 81, at 1523.
[121] *See supra* Parts II.C–II.D.
[122] 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/.

public scrutiny, and because—as discussed above—false claims in such political ads are likely to be spread farther, faster, deeper, and more broadly than true claims in political ads.[123]

F. Measures to Address False Political Advertising and Microtargeting by Social Media Platforms

As of this writing, despite a heightened awareness of the problems caused by microtargeted political advertising and by false political ads, such problems have yet to be effectively addressed via regulation or legislation (at least in the United States). Instead, political advertising on social media, and the regulation of false political speech and microtargeting in particular, is subject to an ad hoc patchwork of voluntary, piecemeal measures recently adopted by the social media platforms themselves. Some of the social media platforms—notably Twitter—are adopting rigorous measures to combating such problems, while others—notably Facebook—have adopted more of a hands-off approach, at least with respect to political ads that constitute "direct speech by politicians."[124] Below I examine the measures undertaken by the social media platforms to address problems caused by the microtargeting of political ads and by false and misleading political ads.

1. Twitter's Regulation of Political Ads

Of the three major social media platforms, Twitter has taken the most aggressive stance with respect to false and misleading political ads by banning political ads altogether. In October 2019, Twitter CEO Jack Dorsey announced that the platform would ban all political advertising.[125] The decision places Twitter in stark contrast with Facebook, which allows political ads and exempts politicians' political ads from its fact-checking program[126] and whose CEO Mark Zuckerberg had stridently defended his company's laissez-faire attitude towards political content moderation on the grounds that this approach

[123] *See* Vosoughi, Roy & Aral, *supra* note 83, at 1146–1151.

[124] *See infra* Part II.F.

[125] Jack Dorsey (@jack), TWITTER (Oct. 30, 2019, 4:05 PM), https://twitter.com/jack/status/1189634360472829952.

[126] Nick Clegg, *Facebook, Elections and Political Speech*, FACEBOOK NEWSROOM (Sept. 24, 2019), https://about.fb.com/news/2019/09/elections-and-political-speech/.

upholds the ideal of free expression.[127] By contrast, Dorsey distinguished Twitter's new policy by explaining that it is not about free expression, but rather about politicians "paying for reach."[128]

Twitter published its policy for implementing its political advertising ban on November 11, 2019, a little less than a year before the 2020 presidential election.[129] Twitter 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."[130] 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.[131] PACs, SuperPACs, candidates, political parties, and elected or appointed government officials are also banned from advertising on Twitter.[132] There are, however, some exemptions. Advertisers that Twitter deems to be news publishers may reference political content so long as the reference does not amount to advocacy.[133]

Twitter's Legal, Policy and Trust & Safety Lead Vijaya Gadde also identified an exemption for "cause-based ads"[134]—ads that "educate, raise awareness, and/or call for people to take action in connection with civic engagement, economic growth,

[127] Cecilia Kang & Mike Isaac, *Defiant Zuckerberg Says Facebook Won't Police Political Speech*, N.Y. TIMES (Oct. 21, 2019), https://www.nytimes.com/2019/10/17/business/zuckerberg-facebook-free-speech.html.

[128] Jack Dorsey (@jack), TWITTER (Oct. 30, 2019, 4:05 PM), https://twitter.com/jack/status/1189634377057067008.

[129] *Political Content*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html (last accessed Sept. 3, 2020).

[130] *Id.*

[131] *Id.*

[132] *Political Content FAQs*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content/political-content-faqs.html (last accessed July 19, 2020).

[133] *Political Content, supra* note 129. Such publishers must have a minimum of 100,000 monthly unique visitors in the United States. *See How to Get Exempted As a News Publisher from the Political Content Policy*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content/news-exemption.html (last accessed July 19, 2020). They must also have a searchable archive, may not be primarily user-generated or aggregated content, and must not be dedicated to a single issue. *Id.*

[134] Vijaya Gadde (@vijaya), TWITTER (Nov. 15, 2019, 1:30 PM), https://twitter.com/vijaya/status/1195408747926917120.

environmental stewardship, or social equity causes."[135] Political organizations, candidates, and politicians may not use such ads, but other groups may.[136] Among other restrictions,[137] cause-based ads may not be microtargeted.[138]

Twitter's allowance of cause-based ads is an apparent response to initial criticism of Twitter's policy. Many users reacted to Twitter's announcement by requesting more precise definitions, including questions about what constitutes a "political" ad[139] and what constitutes an "ad."[140] As yet, it is unclear. Twitter states that for-profit organizations may place "cause ads" if they do not "have the primary goal of driving political, judicial, legislative, or regulatory outcomes" and are "tied to the organization's publicly stated values, principles, and/or beliefs."[141] However, it is not clear at this time how Twitter will interpret the "primary goal" language in its policy.[142]

[135] *Cause-Based Advertising Policy*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/restricted-content-policies/cause-based-advertising.html (last accessed July 19, 2020).

[136] Gadde (@vijaya), *supra* note 134.

[137] Restrictions include certification for caused-based advertisers. *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 accessed July 19, 2020).

[138] *See Cause-Based Advertising Policy*, *supra* note 135.

[139] Aaron Huertas (@aaronhuertas), TWITTER (Oct. 30, 2019, 3:37 PM), https://web.archive.org/web/20191030232518if_/https://twitter.com/aaronhuertas/status/1189672683400761344.

[140] Brad Koenig (@MavsLaker), TWITTER (Oct. 31, 2019, 4:19 PM), https://twitter.com/MavsLaker/status/1190000411559780358.

[141] *Cause-based advertising FAQs*, TWITTER: BUS., https://business.twitter.com/en/help/ads-policies/ads-content-policies/cause-based-advertising/faqs.html (last accessed July 19, 2020).

[142] Some have commented that Sierra Club could promote their causes but not single out politicians or legislation, or that a group could run a gun violence awareness ad but not call for a ban on assault weapons as that would imply a legislative outcome. Sheila Dang & Paresh Dave, *Twitter Tightens Bans on Political Ads and Causes Ahead of 2020 U.S. Election*, REUTERS (Nov. 15, 2019), https://www.reuters.com/article/us-twitter-politics-adban-idUSKBN1XP224. Others have observed the challenges Twitter can expect to face in distinguishing between causes and political outcomes. *See id.* For instance, is an ad about universal healthcare a cause, or is it about a related bill, and how would that be determined? Still others have observed that, if Twitter's misinformation policy is not integrated with its cause-based ads policy, Twitter could still permit inaccurate, but "softer" talking points that don't rise to the level of lobbying, e.g. an anti-minimum wage ad would not be permitted but an inaccurate ad about how a minimum wage law bankrupted a town could conceivably be permitted. Emily Stewart, *Twitter Is Walking into a Minefield with Its Political Ads Ban*, VOX: RECODE (Nov. 15, 2019), https://www.vox.com/recode/2019/11/15/20966908/twitter-political-ad-ban-policies-issue-ads-jack-dorsey.

In addition to prohibiting political ads on its platform, Twitter recently announced measures to combat misinformation in the form of manipulated media like deepfakes and shallow fakes.[143] On February 4, 2020, Twitter announced its new policy on "synthetic and manipulated media," which provides: "[Twitter users] may not deceptively share synthetic or manipulated media that are likely to cause harm." [144] In addition, "[Twitter] may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context."[145] Pursuant to this rule, Twitter will label content that is deceptively altered or fabricated, and will remove content if it impacts public safety or is likely to cause serious harm.[146] Twitter has already shown, on five separate occasions, that it will place warnings on posts from the President that violate its policies, such as its policies on abusive behavior and on misinformation, including manipulated media.[147]

[143] *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. A "deepfake" is a "product of artificial intelligence or machine learning, including deep learning techniques (e.g., a technical deepfake), that merges, combines, replaces, and/or superimposes content onto a video, creating a video that appears authentic." *Community Standards: Manipulated Media*, FACEBOOK, https://m.facebook.com/communitystandards/manipulated_media/ (last accessed Jan. 12, 2021); *see generally* Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753 (2019); Richard L. Hasen, *Deep Fakes, Bots, and Siloed Justices: American Election Law in a "Post-Truth" World*, 64 ST. LOUIS U. L.J. 534 (2020).

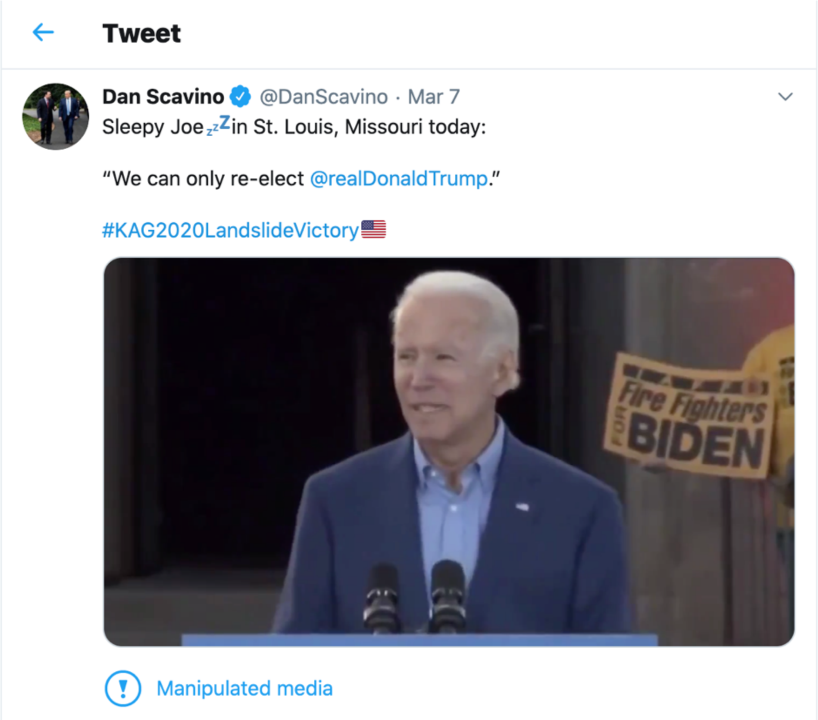
[144] Roth & Achuthan, *supra* note 143.

[145] *Synthetic and Manipulated Media Policy,* TWITTER HELP CENTER, https://help.twitter.com/en/rules-and-policies/manipulated-media (last visited July 19, 2020).

[146] *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 likely to cause harm—may not be shared on Twitter and are subject to removal." *Id.* Additionally, "accounts engaging in repeated or severe violations of this policy may be permanently suspended." *Id.*

[147] 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. Elizabeth Dwoskin, *Twitter's decision to label Trump's Tweets was two years in the making*, WASH. POST (May 29, 2020), https://www.washingtonpost.com/technology/2020/05/29/inside-twitter-trump-label/. As of the time of this writing, Twitter most recently 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), https://www.washingtonpost.com/technology/2020/06/23/twitter-slaps-another-warning-label-trump-tweet-about-force/. Facebook left the post up without a warning. *Id.*

In the first case of Twitter applying its new policy on disinformation through deliberately altered content, Twitter labeled as "manipulated media" an edited video featuring 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.[148] The video had been edited so as to mislead viewers into believing that Biden was actually endorsing Trump.[149]



Tweet

Dan Scavino @DanScavino · Mar 7

Sleepy Joe zzZ in St. Louis, Missouri today:

"We can only re-elect @realDonaldTrump."

#KAG2020LandslideVictory🇺🇸

Manipulated media

Figure Seven: Tweet from Dan Scavino, labeled by Twitter as "Manipulated Media."[150]

In short, Twitter's absolute ban on political ads and its restrictions on manipulated media constitute strong and likely effective measures toward addressing the problems of false and

[148] *See* Ivan Mehta, *Trump's retweet with doctored Biden video earns Twitter's first 'manipulated media' label*, The Next Web (March 9, 2020), https://thenextweb.com/twitter/2020/03/09/trumps-tweet-with-doctored-biden-video-earns-twitters-first-manipulated-media-label/.

[149] *Id.*

[150] Dan Scavino (@DanScavino), Twitter (Mar. 7, 2020, 8:18 PM), https://twitter.com/DanScavino/status/1236461268594294785.

misleading political speech. Some skeptics of the ban, however, have pointed out that the ban will not affect "organic" content or messages from politicians that are shared or retweeted by supporters, and that it could encourage the use of "bots" or paid users to amplify the tweets.[151] In addition, it remains to be seen whether Twitter's carve-out for caused-based ads will provide sufficient opportunities for important speech on topics of civic and social activism.

2a. Facebook's Regulation of Falsity in Political Ads

Facebook is taking a number of steps to combat misinformation on its platform.[152] The company has adopted extensive measures to combat publicly-available misinformation, including by partnering with independent third-party fact checkers to evaluate posts, providing counterspeech in the form of "Related Articles"/"Additional Reporting on This" on topics similar to false or misleading posts, and limiting the distribution of posts from content providers who repeatedly share false news and eliminating their ability to profit.[153] These measures are applicable to political content and political ads, but they are not applicable to posts that are considered "direct speech by a politician."[154] 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."[155] This exception for politicians' content has come under substantial scrutiny in recent months, especially given the highly controversial posts of

[151] AFP, *Twitter Exempts some 'Cause-based' Messages from Political Ad Ban*, FIN. EXPRESS (Nov. 16, 2019), https://www.financialexpress.com/industry/technology/twitter-exempts-some-cause-based-messages-from-political-ad-ban/; *see also* Zack Whittaker, *Twitter Says it Will Restrict Users from Retweeting World Leaders Who Break Its Rules*, TECHCRUNCH (Oct. 15, 2019), https://techcrunch.com/2019/10/15/twitter-world-leaders-break-rules/.

[152] *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 [hereinafter Lyons, *Hard Questions*].

[153] *See* Hunt Allcott et al., *Trends in the Diffusion of Misinformation on Social Media*, app. at 4 (Stan. Inst. for Econ. Pol'y Res., Working Paper No. 18-029, 2018), http://web.stanford.edu/~gentzkow/research/fake-news-trends-appx.pdf (listing in Table 1 all of Facebook's efforts to combat false news).

[154] *Fact-Checking on Facebook: Program Policies*, FACEBOOK BUS. HELP CTR., https://www.facebook.com/business/help/315131736305613?id=673052479947730 (last visited Sept. 4, 2020).

[155] *Id.*

President Trump.[156] Before examining this controversial exception to Facebook's general fact-checking policy for public posts on its platform, I first examine the company's generally-applicable policy itself.

2b. Facebook's General Fact-Checking Policy for Publicly-Available Posts—Excluding the Posts of Politicians

Facebook is continuing to expand the partnership that it began in December 2016 with fact-checkers to evaluate publicly-available content posted on its platform.[157] Through its fact-checking initiatives, Facebook is working with select independent third-party fact checkers, which are certified through the non-partisan International Fact-Checking Network.[158] 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.[159]

Facebook has expanded its fact-checking initiative to include the fact checking of all public, newsworthy Facebook

[156] *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; *see also* 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.

[157] *See* Lyons, *Hard Questions*, *supra* note 152.

[158] *See id.*; *see also Verified Signatories of the IFCN Code of Principles*, POYNTER, https://ifcncodeofprinciples.poynter.org/signatories (last visited Sept. 12, 2020).

[159] *See* Mike Ananny, *Checking in with the Facebook Fact-Checking Partnership,* COLUM. JOURNALISM REV. (Apr. 4, 2018), https://www.cjr.org/tow_center/facebook-fact-checking-partnerships.php; *see also How are independent fact-checkers selected on Facebook?,* FACEBOOK HELP CTR., (explaining process of how a third party becomes a fact-checker for Facebook), https://www.facebook.com/help/1599660546745980?helpref=faq_content (last visited Sept. 29, 2018); *Fact-Checking on Facebook*, FACEBOOK HELP CTR., https://www.facebook.com/help/publisher/182222309230722 (last visited July 19, 2020) (providing an overview of Facebook's fact-checking program). 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. *See* Aaron Rupar, *Facebook's Controversial Fact-checking Partnership with a Daily Caller-funded Website, Explained*, VOX (May 6, 2019), https://www.vox.com/2019/5/2/18522758/facebook-fact-checking-partnership-daily-caller.

posts, including links, articles, photos, and videos.[160] The fact-checking process on Facebook also applies to political advertisements unless those advertisements (or other posts) constitute the speech of politicians.[161] As Facebook explains:

> We don't believe . . . 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. That's why Facebook exempts politicians from our third-party fact-checking program . . . . This means that we will not send organic content or ads from politicians to our third-party fact-checking partners for review . . . . [W]e do not submit speech by politicians to our independent fact-checkers, and we generally allow it on the platform even when it would otherwise breach our normal content rules.[162]

This conspicuous exception to Facebook's fact-checking process has major ramifications for the political process, and has subjected Facebook to substantial criticism in recent months. Below, I first examine Facebook's fact-checking process generally, and then turn to the exception to this process for posts made by elected officials (including their political advertisements).

Facebook's fact-checking process can be initiated by Facebook users who flag a post as being potentially false.[163] 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

[160] *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.

[161] "If the 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." *Fact-Checking on Facebook: Program Policies*, *supra* note 154.

[162] *Facebook, Elections, and Political Speech*, FACEBOOK, https://about.fb.com/news/2019/09/elections-and-political-speech/ (last visited Jan. 12, 2021).

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

journalist, or, as is most commonly the case, by Facebook's machine learning algorithms.[164] For a user to flag a post as potentially false, a user must click "•••" next to the post he or she wishes to flag as false, then click "Report post," then click "False News."[165]

Once a post is flagged by a user as a potential false news story, it is submitted for evaluation to a third-party independent fact-checker.[166] While the process of evaluating posts in the past was triggered only by user flagging, Facebook now incorporates other ways of triggering such evaluation, including by providing its independent fact-checkers with the authority to proactively identify posts to review[167] as well as by using machine learning to identify potentially false posts.[168] 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.[169]

Once a third-party fact-checker has determined that a post is false, Facebook then initiates several steps. First, Facebook deprioritizes false posts in users' News Feeds, i.e. the constantly updating list of stories in the middle of a user's home page (including status updates, photos, videos, links, app activity, and likes), such that future views of each false post will be reduced by an average of eighty percent.[170] Second, Facebook commissions a fact-checker to write a "Related Article" setting forth truthful information about the subject of the false post and the reasons why the fact-checker rated the post as false.[171] Such content is

[164] *See Fact-Checking on Facebook: Program Policies*, *supra* note 154.

[165] *See How do I Mark a Facebook Post as False News?*, *supra* note 163.

[166] *See* Lyons, *Hard Questions*, *supra* note 152 ("[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.").

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

[168] *See* Dan Zigmond, *Machine Learning, Fact-Checkers and the Fight Against False News*, FACEBOOK NEWSROOM (Apr. 8, 2018), https://about.fb.com/news/2018/04/inside-feed-misinformation-zigmond.

[169] *Fact-Checking on Facebook: Facebook's Enforcement of Fact-Checker Ratings*, FACEBOOK BUS. HELP CTR., https://www.facebook.com/business/help/341102040382165?id=673052479947730 (last accessed August 28, 2020).

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

[171] *See* Tessa Lyons, *Replacing Disputed Flags with Related Articles*, FACEBOOK NEWSROOM (Dec. 20, 2017), https://newsroom.fb.com/news/2017/12/news-feed-

then displayed in conjunction with the false post on the same subject.[172] While Facebook formerly flagged false news sites with a "Disputed" flag, the company changed its approach in response to research suggesting that such flags may actually entrench beliefs in the disputed posts.[173] Facebook now provides "Related Articles" in conjunction with false news stories, which apparently does not result in similar entrenchment.[174] 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," as will users who earlier shared the false post,[175] as in the example below (setting forth Facebook and Instagram's flags).

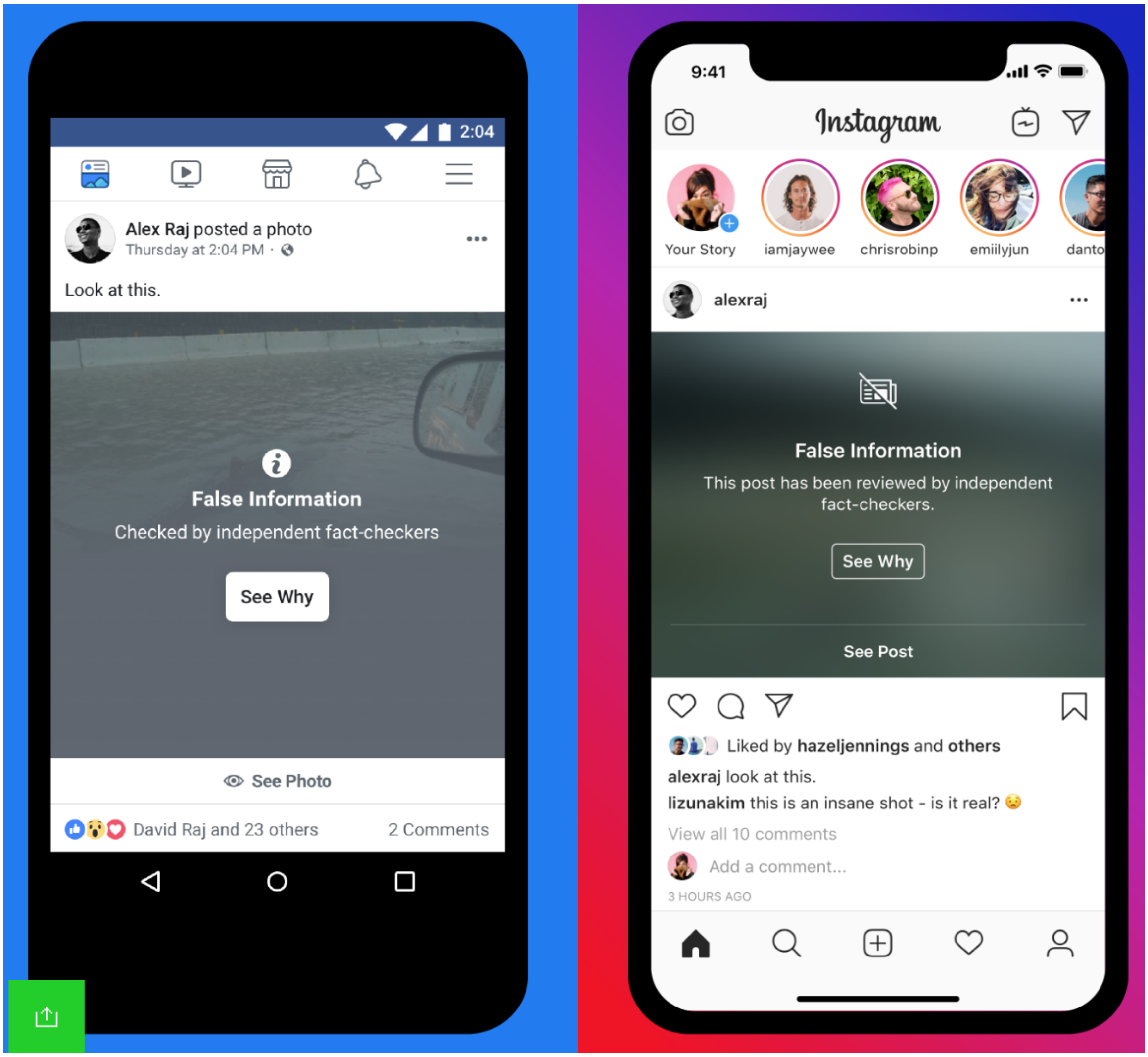
fyi-updates-in-our-fight-against-misinformation [hereinafter Lyons, *Replacing Disputed Flags*].

[172] *See id.*; *see also* Geoffrey A. Fowler, *I fell for Facebook fake news. Here's why millions of you did, too.*, WASH. POST (Oct. 18, 2018), https://www.washingtonpost.com/technology/2018/10/18/i-fell-facebook-fake-news-heres-why-millions-you-did-too/ (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, *Replacing Disputed Flags*, *supra* note 171; Sara Su, *New Test with Related Articles*, FACEBOOK NEWSROOM (Apr. 25, 2017), https://newsroom.fb.com/news/2017/04/news-feed-fyi-new-test-with-related-articles.

[173] *See* Lyons, *Replacing Disputed Flags*, *supra* note 171.

[174] *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.").

[175] *See id.*



2:04

Alex Raj posted a photo
Thursday at 2:04 PM ·

Look at this.

False Information
Checked by independent fact-checkers

See Why

See Photo

David Raj and 23 others 2 Comments

9:41

Instagram

Your Story iamjaywee chrisrobinp emiilyjun danto

alexraj

False Information
This post has been reviewed by independent fact-checkers.

See Why

See Post

Liked by **hazeljennings** and **others**

alexraj look at this.

lizunakim this is an insane shot - is it real? 😮

View all 10 comments

Add a comment...

3 HOURS AGO

Figure Eight: Examples of false information labels on Instagram.[176]

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."[177] In addition, Facebook will now post more prominent fact-checking labels as interstitial warnings atop photos and videos on Facebook (and Instagram) that were fact-checked as false.

[176] 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/.

[177] *How is Facebook Addressing False Information through Independent Fact-checkers*, FACEBOOK BUS. HELP CTR., https://www.facebook.com/help/1952307158131536 (last visited July 21, 2020).

Third, content providers—i.e., Facebook pages and domains—that repeatedly publish and/or share false posts will have their ability to monetize and advertise reduced and ultimately disabled by Facebook unless and until they issue corrections or successfully dispute fact-checkers' determination that their posts are false.[178] Facebook's nimble and extensive efforts to combat publicly available misinformation with labeling and counterspeech are commendable and should be expanded to include the direct speech of politicians as well, as I examine below.

2c. Facebook's Policies Regarding False Political Ads

With respect to false political ads, Facebook's policy is complex. Although Facebook has implemented extensive measures with respect to false posts generally (described above), this false news policy does not apply to "direct speech" by politicians.[179] Accordingly, Facebook's general false news policy, composed of the fact-checking process described above, has an exception for "direct speech" by politicians, such that direct speech by politicians is not run through Facebook's external fact checking process.[180] Facebook provides 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.[181]

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

[179] *Fact-Checking on Facebook: Program Policies*, *supra* note 154.

[180] *Id.*

[181] Nick Clegg, *Facebook, Elections and Political Speech*, FACEBOOK NEWSROOM (Sep. 24, 2019), https://about.fb.com/news/2019/09/elections-and-political-speech/.

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.[182] Facebook further justifies its policies as follows: "In a democracy, people should decide what is credible, not tech companies . . . . That's why—like other [I]nternet platforms and broadcasters—we don't fact check ads from politicians."[183] Facebook also defends its decision by adverting to the importance of political ads to challengers and local candidates: "Given the sensitivity around political ads, we have considered whether we should ban them altogether . . . But political ads are important for local candidates, up-and-coming challengers, and advocacy groups that use our platform to reach voters and their communities."[184]

As a result, political speech and political posts and campaign ads made by politicians themselves operate in a separate system on Facebook. While ordinary users who publicly post false content may face consequences, including being banned from Facebook,[185] elected officials are exempt.

Facebook's policies came into sharp focus in October 2019, when President Donald Trump's reelection campaign began running an ad that was proven to be false about then former Vice President Joe Biden on Facebook.[186] The Trump Campaign released a 30-second video ad accusing then former Vice President Biden of promising Ukraine $1 billion in aid in exchange for firing a prosecutor who was investigating a

Facebook will not fact check political ads from candidates, but it does evaluate and fact-check political ads from political advocacy groups or political action committees. *See* David Klepper, *Facebook Clarifies Zuckerberg Remarks on False Political Ads*, AP NEWS (Oct. 24, 2019), https://apnews.com/64fe06acd28145f5913d6f815bec36a2.

[182] *Id.*

[183] Klepper, *supra* note 181.

[184] *Id.*

[185] *See* Lyons, *Hard Questions*, *supra* note 152 (explaining that Facebook takes action against accounts that share false news, including removing accounts that share false news that also violates other Facebook Community Standards).

[186] 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/.

company with ties to Biden's son, Hunter Biden.[187] The Biden Campaign asked Facebook to take down the ad, but Facebook refused.[188] 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."[189] Accordingly, the false Trump Campaign ad on Biden remained on Facebook, garnering at least 4.6 million views.[190]

Former presidential candidate Senator Elizabeth Warren, who has a history of locking horns with Facebook and with big tech in general,[191] took particular aim at Facebook's policy towards political ads by placing an intentionally false ad on the platform in October 2019.[192] Warren's ad declared that "Mark Zuckerberg and Facebook just endorsed Donald Trump for re-

[187] Grynbaum & Hsu, *supra* note 156.

[188] Kang, *supra* note 156.

[189] *Id.*

[190] Jeremy B. Merrill, *While everyone was looking at Facebook, Trump's false Biden ad appeared more often on YouTube*, QUARTZ (Nov. 1, 2019), https://qz.com/1739780/trumps-biden-ad-appeared-more-often-on-youtube-than-on-facebook/.

[191] *See* Elizabeth Warren, (@TeamWarren), *Here's how we can break up Big Tech*, MEDIUM (Mar. 8, 2019), https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c. After announcing her ambition to break up big tech companies, Warren took out ads on Facebook that denounced Facebook itself as well as Amazon and Google for their "vast power over our economy and our democracy." Cristiano Lima, *Facebook backtracks after removing Warren ads calling for Facebook breakup*, POLITICO (Mar. 11, 2019), https://www.politico.com/story/2019/03/11/facebook-removes-elizabeth-warren-ads-1216757. Facebook initially removed the ads, apparently because they contained an unauthorized reproduction of Facebook's logo, but soon after, the company reversed course and restored them "[i]n the interest of allowing robust debate." Isaac Stanley-Becker & Tony Romm, *Facebook Deletes, and then Restores, Elizabeth Warren's Ads Criticizing the Platform, Drawing her Rebuke*, WASH. POST (Mar. 12, 2019), https://www.washingtonpost.com/nation/2019/03/12/facebook-deletes-then-restores-elizabeth-warrens-ads-criticizing-platform-drawing-her-rebuke/. Warren meanwhile warned of the danger of a "social media marketplace" that is "dominated by a single censor." Elizabeth Warren (@ewarren), TWITTER (Mar. 11, 2019, 7:59 PM), https://twitter.com/ewarren/status/1105256905058979841; Elizabeth Warren, *Elizabeth's plan: Break up the big tech companies*, FACEBOOK (Mar. 8, 2019), https://www.facebook.com/ElizabethWarren/videos/396777104233421/.

[192] Brian Fung, *Elizabeth Warren Targets Facebook's Ad Policy -- with a Facebook Ad*, CNN (Oct. 12, 2019, 12:11 PM), https://www.cnn.com/2019/10/11/politics/elizabeth-warren-facebook-ad/index.html.

election."[193] She explained that the ad was a test to see "just how far" Facebook's policy went and accused Facebook of becoming a "disinformation-for-profit machine."[194] Adhering to its policy of refusing to fact-check direct speech by politicians, Facebook declined to remove Warren's intentionally (and provocatively) false ad, stating "if Senator Warren wants to say things she knows to be untrue, we believe Facebook should not be in the position of censoring that speech."[195]

In addition, Facebook employees recently rose up in strong opposition to Facebook's 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).[196] The particular flashpoint most recently at issue involved violent speech, not misinformation, in the form of Donald Trump's May 2020 post following the murder of George Floyd and the ensuing demonstrations.[197] Trump threatened to deploy the military in Minneapolis to "bring the City under control" and infamously stated "when the looting starts, the shooting starts."[198]

[193] Elizabeth Warren (@ewarren), TWITTER (Oct. 12, 2019, 10:01 AM), https://twitter.com/ewarren/status/1183019897804197888?s=20.

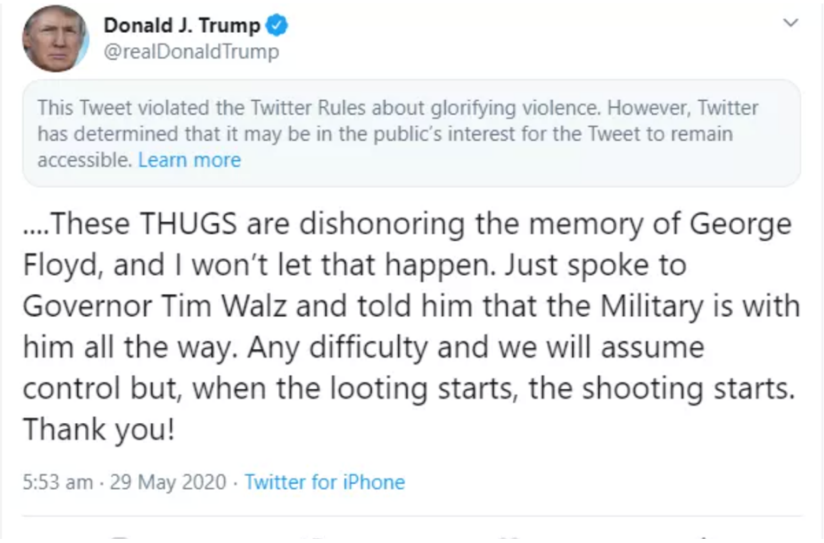
[194] *Id.*

[195] Fung, *supra* note 192.

[196] *See* Megan Rose Dickey & Taylor Hatmaker, *Facebook Employees Stage Virtual Walkout in Protest of Company's Stance on Trump Posts*, TECHCRUNCH (June 1, 2020, 1:01 PM), https://techcrunch.com/2020/06/01/facebook-employees-stage-virtual-walkout-in-protest-of-companys-stance-on-trump-posts/; *see also* Rachel Siegel & Elizabeth Dwoskin, *Facebook Employees Blast Zuckerberg's Hands-off Response to Trump Posts as Protests Grip Nation*, WASH. POST (June 1, 2020, 8:04 PM), https://www.washingtonpost.com/business/2020/06/01/facebook-zuckerberg-donation-trump/.

[197] Dickey & Hatmaker, *supra* note 196.

[198] *Id.*



Donald J. Trump
@realDonaldTrump

This Tweet violated the Twitter Rules about glorifying violence. However, Twitter has determined that it may be in the public's interest for the Tweet to remain accessible. Learn more

....These THUGS are dishonoring the memory of George Floyd, and I won't let that happen. Just spoke to Governor Tim Walz and told him that the Military is with him all the way. Any difficulty and we will assume control but, when the looting starts, the shooting starts. Thank you!

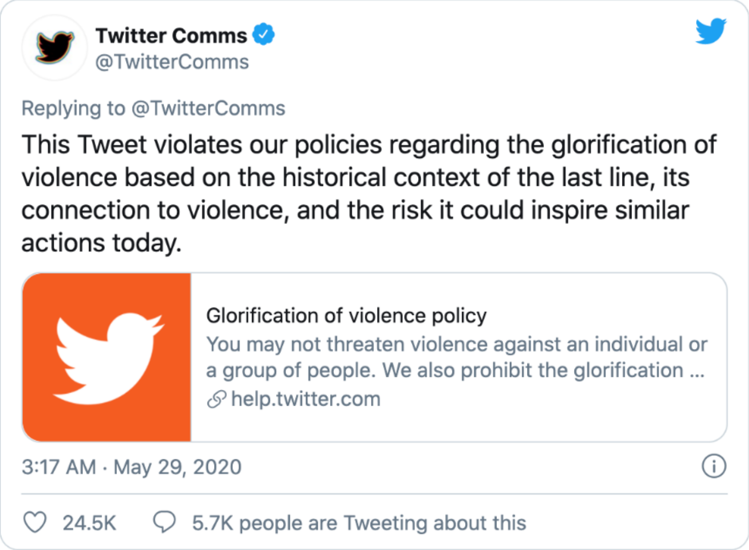
5:53 am · 29 May 2020 · Twitter for iPhone

Figure Nine: Tweet from Donald Trump following the murder of George Floyd in May 2020.[199]

President Trump made this post across multiple platforms.[200] 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),

[199] Donald J. Trump (@realDonaldTrump), TWITTER (May 29, 2020, 12:53 AM), https://twitter.com/realDonaldTrump/status/1266231100780744704.

[200] *See id.*



Twitter Comms @TwitterComms

Replying to @TwitterComms

This Tweet violates our policies regarding the glorification of violence based on the historical context of the last line, its connection to violence, and the risk it could inspire similar actions today.

Glorification of violence policy
You may not threaten violence against an individual or a group of people. We also prohibit the glorification ...
help.twitter.com

3:17 AM · May 29, 2020

24.5K 5.7K people are Tweeting about this

Figure Ten: Twitter's explanation that Donald Trump's tweet following the murder of George Floyd violated the platform's rules against glorifying violence.[201]

Facebook took no action.[202] 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."[203] Zuckerberg explained further that "we read [Trump's post] as a warning about state action, and we think people need to know if the government is planning to deploy force."[204] Some of Facebook's employees, however, were extremely dissatisfied by the company's response, resulting in "intense debate" on Facebook's internal employee messaging system about the company's laissez-faire policies regarding politicians' posts.[205] In response, Zuckerberg hosted an internal town-hall to explain his and the company's rationale for

[201] Twitter Comms (@TwitterComms), TWITTER (May 29, 2020, 3:17 AM), https://twitter.com/TwitterComms/status/1266267447838949378.

[202] Brian Stelter & Donie O'Sullivan, *Trump Tweets Threat That 'Looting' Will Lead to 'Shooting.' Twitter Put a Warning Label on It*, CNN BUSINESS (May 29, 2020, 10:40 AM) (screenshot included), https://www.cnn.com/2020/05/29/tech/trump-twitter-minneapolis/index.html.

[203] Mark Zuckerberg, FACEBOOK (May 29, 2020, 4:19 PM), https://www.facebook.com/zuck/posts/10111961824369871.

[204] *Id.*

[205] Siegel & Dwoskin, *supra* note 196.

inaction.[206] Facebook ultimately retreated from its non-interventionist stance towards Donald Trump and his campaign, at least with respect to its hate speech content regulation, as it removed a Trump Campaign page ad because it used a hate symbol.[207] However, many companies felt Facebook still had not gone far enough and joined a growing advertising boycott to pressure the platform to take more aggressive action against the hate speech and misinformation being spread by political figures such as President Trump.[208] Facebook responded by announcing "that it would remove posts [from political leaders] that incite

[206] Elizabeth Dwoskin & Nitasha Tiku, *Facebook Employees Said They were 'Caught in an Abusive Relationship' with Trump as Internal Debates Raged*, WASH. POST (June 5, 2020, 4:37 PM), https://www.washingtonpost.com/technology/2020/06/05/facebook-zuckerberg-trump/.

[207] Isaac Stanley-Becker, *Facebook Removes Trump Ads with Symbol Once Used by Nazis to Designate Political Prisoners*, WASH. POST (June 18, 2020), https://www.washingtonpost.com/politics/2020/06/18/trump-campaign-runs-ads-with-marking-once-used-by-nazis-designate-political-prisoners/. Days later when a Trump-affiliated campaign page posted an advertisement denouncing "dangerous MOBS of far-left groups . . . causing absolute mayhem" accompanied by an image of a downward facing red 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 symbols" outside of a condemnatory context or as an object for discussion. *Id.* Zuckerberg has also since announced that Facebook will begin labeling "newsworthy content." Mark Zuckerberg, FACEBOOK (June 26, 2020, 11:25 AM), https://www.facebook.com/zuck/posts/10112048980882521. Occasionally, he explains, "we leave up content that would otherwise violate our policies if the public interest value outweighs the risk of harm." *Id.* Now, Facebook will append a notification that the content violates Facebook's policy but remains so that people can engage with and discuss it. *Id.* Facebook will also further restrict content that can be included in paid advertisements. *Id.* Ads that claim people from "a specific race, ethnicity, national origin, religious affiliation, caste, sexual orientation, gender identity or immigration status are a threat to the physical safety, health or survival of others" are now prohibited when they were not before, and Facebook also intends to "better protect immigrants, migrants, refugees and asylum seekers from ads suggesting these groups are inferior or expressing contempt, dismissal or disgust directed at them." *Id.*

[208] *All the Companies Quitting Facebook*, N.Y. TIMES: DEALBOOK NEWSLETTER (July 7, 2020) ("Marketers are expressing unease with how [Facebook] handles misinformation and hate speech, including its permissive approach to problematic posts by President Trump."), https://www.nytimes.com/2020/06/29/business/dealbook/facebook-boycott-ads.html. For a list of companies boycotting, see Tiffany Hsu & Gillian Friedman, *CVS, Dunkin', Lego: The Brands Pulling Ads from Facebook Over Hate Speech*, N.Y. TIMES (July 7. 2020), https://www.nytimes.com/2020/06/26/business/media/Facebook-advertising-boycott.html; Allen Kim & Brian Fung, *Facebook boycott: View the list of companies pulling ads*, CNN (July 2, 2020, 6:05 PM), https://www.cnn.com/2020/06/28/business/facebook-ad-boycott-list/index.html.

violence or attempt to suppress voting . . . [and] affix labels on posts that violate hate speech prohibitions."[209]

Facebook's policies also drew sharp criticism from civil rights experts, who conducted an extensive, independent two-year civil rights audit of Facebook's content regulation policies and their implementation.[210] The experts' 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.[211] The civil rights experts strongly criticized Facebook'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:

> We have grave concerns that the combination of the company'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

[209] Craig Timberg & Elizabeth Dwoskin, *Silicon Valley is getting tougher on Trump and his supporters over hate speech and disinformation,* WASH. POST (July 10, 2020), https://www.washingtonpost.com/technology/2020/07/10/hate-speech-trump-tech/. Twitch recently suspended President Trump's account and Reddit closed a long-controversial forum named after the President (this same forum helped to popularize the dangerous Pizzagate false conspiracy theory). *Id.* Reddit's action may have been in response to "employee" concerns as well, as it came after an open letter written by hundreds of volunteer moderators chastised Reddit's leadership for the proliferation of hateful speech, calling it the company's "most glaring problem." *Id.*

[210] *See* MURPHY ET AL., FACEBOOK'S CIVIL RIGHTS AUDIT – FINAL REPORT, *supra* note 18.

[211] *Id.* at 37–38.

> communities to suppress the votes of groups based on their race or other characteristics. With only months left before a major election, this is deeply troublesome as misinformation, sowing racial division and calls for violence near elections can do great damage to our democracy.[212]

The concerns of the civil rights experts turned out to be well-founded. The company's reticence to take decisive action regarding the hateful and dangerous rhetoric of politicians has indeed brought about great damage to our democracy.

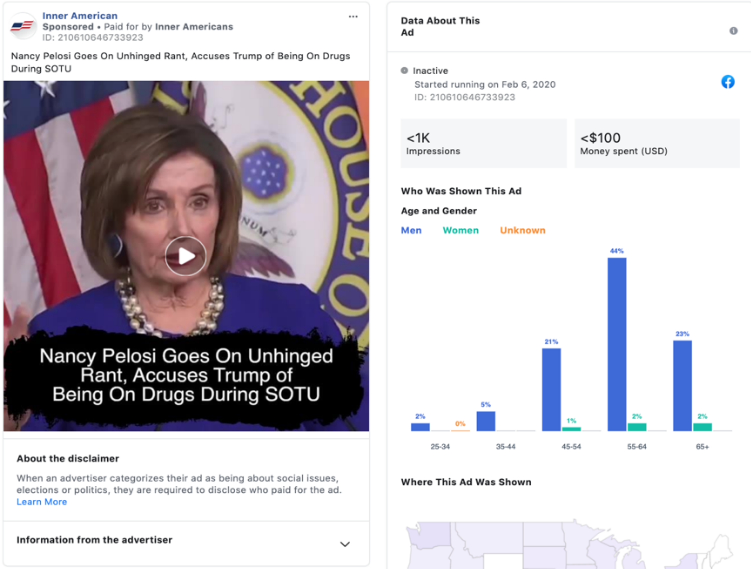
2d. Facebook's Transparency and Disclosure Requirements Regarding Political/Electioneering Advertisements

Facebook recently implemented a Political Advertising Policy that requires, first, that every election-related and issue advertisement made available on Facebook to users in the United States be clearly labeled as a "Political Ad" and include a "Paid for by" disclosure, with the name of the individual or organization who paid for the advertisement at the top.[213] Second, under the Policy, Facebook has committed to collecting and maintaining a publicly available archive of political advertisements as part of its Ad Library, which provides information about "the campaign budget associated with an individual ad and how many people saw it—including their age, location, and gender."[214] See example below.

[212] *Id.* at 10.

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

[214] 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/. *See also* MURPHY, *supra* note 18, 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.").



Inner American
Sponsored • Paid for by Inner Americans
ID: 210610646733923

Nancy Pelosi Goes On Unhinged Rant, Accuses Trump of Being On Drugs During SOTU

Nancy Pelosi Goes On Unhinged Rant, Accuses Trump of Being On Drugs During SOTU

About the disclaimer

When an advertiser categorizes their ad as being about social issues, elections or politics, they are required to disclose who paid for the ad. Learn More

Information from the advertiser

Data About This Ad

Inactive
Started running on Feb 6, 2020
ID: 210610646733923

<1K
Impressions

<$100
Money spent (USD)

Who Was Shown This Ad

Age and Gender

Where This Ad Was Shown

Figure Eleven: Example from Facebook's Ad Library.[215]

Facebook has also recently updated its Ad Library to increase transparency and provide more useful data—including by permitting users to search for and filter ads based on the estimated audience size—which enables researchers, advocates, and the public to identify and study micro-targeted ads.[216] Finally, under the Policy, Facebook will prohibit foreign entities from purchasing political ads directed at U.S. audiences.[217] Facebook enforces this by mailing prospective political advertisers a postcard to a U.S. address to verify U.S. residency.[218] If a prospective purchaser of a political ad is not verified under this process, he or she will not be able to purchase a political ad on Facebook.[219] Commenting on the recently implemented Political Advertising Policy, Facebook's CEO Mark Zuckerberg explained, "These changes won't fix everything, but they will make it a lot harder for anyone to do what the Russians did during the 2016 election and use fake

[215] 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).

[216] *See* MURPHY, *supra* note 18, at 35–37.

[217] *Get Authorized to Run Ads About Social Issues, Elections or Politics*, FACEBOOK BUS. HELP CTR., https://www.facebook.com/business/help/208949576550051?id=288762101909005&recommended_by=241608613261133 (last visited Oct. 16, 2020).

[218] *See id.*

[219] *See id.*

accounts and pages to run ads."[220] Facebook's recently implemented measures imposing disclosure requirements on political ads and limiting foreign entities from purchasing political ads go beyond those that are encompassed in the proposed Honest Ads Act, and may at least be moderately successful in preventing the type of foreign interference in U.S. elections that occurred in 2016.

The challenges of microtargeting and foreign influence have further complicated Facebook's efforts to mitigate the harms of political misinformation and disinformation by its users and especially its advertisers, and the revelations surrounding the election of 2016 serve as a potent reminder of the potential dangers of failing to do so. Facebook was slow to get started in taking responsibility for what happens on its platform, but now the platform seems to be trending in the right direction in regulating political advertising and other controversial political speech on its platform, in the absence of actual government regulation.

3a. Google's Measures to Address Microtargeting of Political Ads

Google recently amended its rules governing the practice of microtargeting of political advertisements.[221] While Google maintains that it has never offered "granular microtargeting" of election ads, in November 2019, Google officially 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).[222] Political advertisers can also use contextual targeting, which enables them to serve users with ads according to the content that users are accessing.[223] Google claims this approach aligns it with industry practice in television, radio and print media.[224] Google's policy on microtargeting took effect in the European Union at the end

[220] Josh Constine, *Facebook and Instagram Launch US Political Ad Labeling and Archive*, TECHCRUNCH (May 24, 2018, 2:01 PM), https://techcrunch.com/2018/05/24/facebook-political-ad-archive/.
[221] *See* Scott Spencer, *An Update on Our Political Ads Policy*, GOOGLE BLOG: THE KEYWORD (Nov. 20, 2019), https://blog.google/technology/ads/update-our-political-ads-policy/.
[222] *Id.*
[223] *Id.*
[224] *Id.*

of 2019, and became effective worldwide (including in the United States) in January 2020.[225]

Accordingly, 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.[226] All other types of targeting are not allowed for use in election ads, including the use of Google's powerful Audience Targeting products,[227] Remarketing,[228] Customer Match,[229] and Geographic Radius Targeting.[230] 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.[231] In an email to political campaigns, Google outlined these new rules, explaining that 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.[232] Further, political campaigns can no longer upload their own lists of people to whom they wish to show ads.[233] 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.[234]

[225] *See* Rachel Sandler, *Google Limits Microtargeting for Paid Political Ads*, FORBES (Nov. 20, 2019, 8:22 PM), https://www.forbes.com/sites/rachelsandler/2019/11/20/google-limits-microtargeting-for-paid-political-ads/#55c667fd51ec.

[226] *Election Ads in the United States*, *Political Content*, GOOGLE ADVERT. POLICIES HELP, https://support.google.com/adspolicy/answer/6014595 (last accessed July 21, 2020) [hereinafter *Election Ads in the United States*, *Political Content*].

[227] *About Audience Targeting*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/2497941 (last accessed: July 21, 2020).

[228] *About Remarketing*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/2453998 (last accessed: July 21, 2020).

[229] *About Customer Match*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/6379332 (last accessed: July 21, 2020).

[230] *Target Ads to Geographic Locations*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/1722043 (last accessed: July 21, 2020).

[231] Spencer, *supra* note 221.

[232] Jenna Lowenstein (@just_jenna), TWITTER (Nov. 20, 2019, 6:54 PM), https://twitter.com/just_jenna/status/1197302201938567168; *see also Election Ads in the United States*, *Political Content*, *supra* note 226.

[233] *Election Ads in the United States*, *Political Content*, *supra* note 226.

[234] *Id.*

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.[235] 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.[236] Google's sophisticated targeting tools also have allowed advertisers to target or exclude audiences according to demographic data such as age, gender, household income, homeownership, and the like.[237] General advertisers may also target users who have previously interacted with their site[238] or by submitting previously collected customer data to re-engage with the same group or expand to similar audiences.[239] These sophisticated targeting tools are now unavailable to political advertisers.[240]

One of the greatest challenges Google faces in implementing its policy restricting the use of microtargeting by political advertisers is how to meaningfully and accurately define political/election advertising. With respect to the United States, Google currently defines election ads as those that feature:

1. A current officeholder or candidate for an elected federal office (including federal offices such as that of the President or Vice President of the United States, members of the United States House of Representatives or United States Senate).

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

[236] *About Custom Audiences*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/9805516?hl=en&ref_topic=3122880 (last accessed Oct. 16, 2020).

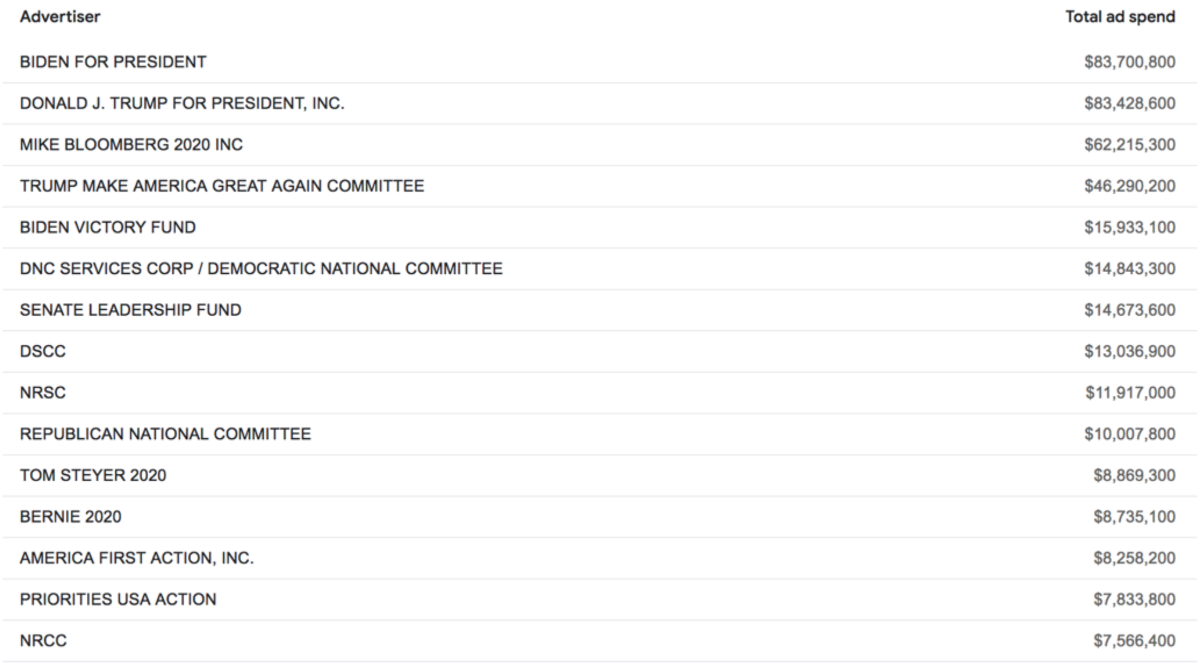
[237] *About Demographic Targeting*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/2580383?hl=en&ref_topic=3122881 (last accessed July 21, 2020).

[238] *Remarketing: Reach People Who Visited Your Site or App*, GOOGLE ADS HELP, https://support.google.com/google-ads/topic/3122874?hl=en&ref_topic=3121935 (last accessed July 21, 2020).

[239] *About Customer Match*, GOOGLE ADS HELP, https://support.google.com/google-ads/answer/6379332?hl=en&ref_topic=6296507 (last accessed July 21, 2020).

[240] *Election Ads in the United States*, *Political Content*, *supra* note 226.

2. A current officeholder or candidate for a state-level elected office, such as Governor, Secretary of State, or member of a state legislature.
3. A federal or state level political party.
4. A state-level ballot measure, initiative, or proposition that has qualified for the ballot in its state.[241]



Advertiser	Total ad spend
BIDEN FOR PRESIDENT	$83,700,800
DONALD J. TRUMP FOR PRESIDENT, INC.	$83,428,600
MIKE BLOOMBERG 2020 INC	$62,215,300
TRUMP MAKE AMERICA GREAT AGAIN COMMITTEE	$46,290,200
BIDEN VICTORY FUND	$15,933,100
DNC SERVICES CORP / DEMOCRATIC NATIONAL COMMITTEE	$14,843,300
SENATE LEADERSHIP FUND	$14,673,600
DSCC	$13,036,900
NRSC	$11,917,000
REPUBLICAN NATIONAL COMMITTEE	$10,007,800
TOM STEYER 2020	$8,869,300
BERNIE 2020	$8,735,100
AMERICA FIRST ACTION, INC.	$8,258,200
PRIORITIES USA ACTION	$7,833,800
NRCC	$7,566,400

Figure Twelve: Google's top political advertisers in the United States.[242]

Yet, few election ads as they are popularly understood are likely to be so specific. For example, "issue ads" funded by Super-PACs may not specifically "advocate the election or defeat of a clearly identified federal candidate,"[243] yet such outside spending makes up the vast majority of political

[241] *Id.*

[242] For data on Google's top political advertisers in the United States since May 31, 2018, see *Political Advertising in the United States,* GOOGLE TRANSPARENCY REP. https://transparencyreport.google.com/political-ads/region/US (last visited Feb. 22, 2021).

[243] *Advertising and Disclaimers*, FED. ELECTION COMM'N, *supra* note 96.

advertising.[244] Thus, Google's definition of election ads may turn out to be substantially underinclusive and ineffective.

Google has also implemented a host of procedural requirements for political advertisers. Advertisers who wish to purchase and run election ads[245] or use political affiliation in personalized advertising[246] in the United States must go through a verification process, which is required for all ad formats/extensions, and all personalized ads features.[247] Political advertisers must provide a Federal Election Commission (FEC) ID and either an Employer Identification Number (EIN) (for organizations) or Social Security Number (for individuals).[248] Google collects such data and makes available a transparency report on political ad spending by each advertiser/campaign.[249] The transparency report lists top advertisers and the amount of political ad spending by each advertiser.[250] A recent transparency report (as of June 6, 2020) provides this list of top political ad spending since May 31, 2018.[251]

3b. Google's Regulation of Falsity and Misleading Content in Political Ads

Google also recently revised its rules about truth-in-advertising to prohibit ads with "demonstrably false claims that could significantly undermine participation or trust" in

[244] *2020 Outside Spending, by Race*, OPENSECRETS.ORG, https://www.opensecrets.org/outsidespending/summ.php?disp=R (last updated Sept. 4, 2020).

[245] *Election Ads in the United States*, *Political Content*, *supra* note 226.

[246] *Political Affiliation in Personalized Advertising*, *Political Content*, GOOGLE ADVERT. POLICIES HELP, https://support.google.com/adspolicy/answer/143465?#533 (last accessed July 21, 2020).

[247] *About Verification for Election Advertising in the United States*, GOOGLE ADVERT. POLICIES HELP, https://support.google.com/adspolicy/answer/9002729?hl=en (last accessed Oct. 16, 2020).

[248] *Id.*; *see also Apply for Verification for Election Advertising in the United States*, GOOGLE DISPLAY & VIDEO 360 HELP, https://support.google.com/displayvideo/answer/9014141?hl=en (last accessed Oct. 16, 2020).

[249] *See, e.g.*, *Political Advertising in the United States*, *supra* note 242.

[250] *Id.*

[251] *Id.*

elections.[252] Google has stated, however, that by reframing these truth-in-advertising rules, it does not intend to appoint itself as the arbiter of truth in politics.[253] Google explains that since "no one can sensibly adjudicate every political claim, counterclaim, and insinuation," it will focus its efforts on claims that are something more than generic falsehood or exaggeration.[254] It will not take comprehensive action against every misleading political ad but will do so for "clear violations."[255] That line will likely be difficult to define and maintain. In its announcement, Google gives the example of "deep fakes" as the type of content that it will now remove.[256] These are addressed by Google's policy prohibiting "manipulating media to deceive, defraud, or mislead others."[257] The example the company provides is "deceptively doctoring media related to politics, social issues, or matters of public concern."[258] Google has also released an open-source database containing 3,000 manipulated videos in order to help identify and target deepfakes.[259]

It is as yet unclear what falls within the category of demonstrably false political ads according to Google,[260] but a few examples provide some guidance. When YouTube CEO Susan Wojcicki was asked whether YouTube would remove President Trump's advertisement (which he placed on Facebook) falsely accusing Joe Biden of corruptly sheltering his son from a Ukrainian investigation through bribery, Wojcicki explained that this ad "would not be a violation of our policies" because

[252] Spencer, *supra* note 221; *Misrepresentation*, GOOGLE ADVERT. POLICIES HELP, https://support.google.com/adspolicy/answer/6020955?hl=en (last accessed July 21, 2020).

[253] *See* Spencer, *supra* note 221.

[254] *Id.*

[255] *Id.*

[256] *Id.*

[257] *Misrepresentation*, *supra* note 252.

[258] *Id.*

[259] Karen Hao, *Google Has Released a Giant Database of Deepfakes to Help Fight Deepfakes*, MIT TECH. REV. (Sept. 25, 2019), https://www.technologyreview.com/f/614426/google-has-released-a-giant-database-of-deepfakes-to-help-fight-deepfakes/; *see also* Nick Dufour & Andrew Gully, *Contributing Data to Deepfake Detection Research*, GOOGLE AI BLOG (Sept. 24, 2019), https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[260] Google and YouTube have removed over 300 Trump ads in the last half of 2019, but the archive in which removed ads are listed does not indicate why specific ads were removed. Shachar Bar-On & Natalie Jimenez Peel, *300+ Trump Ads Taken Down by Google, YouTube*, CBS NEWS: 60 MINUTES OVERTIME (Dec. 1, 2019), https://www.cbsnews.com/news/300-trump-ads-taken-down-by-google-youtube-60-minutes-2019-12-01/.

"politicians are always accusing their opponents of lying."[261] On the other hand, Wojcicki cited the (now infamous) video that showed Nancy Pelosi speaking at an artificially reduced rate, which made Pelosi appear to be drunk.[262] Wojcicki noted that that video was removed "very fast" because "it's not okay to have technically manipulated content that would be misleading."[263]

With respect to manipulated media in particular, YouTube has adopted specifically applicable policies.[264] Its deceptive practices policies state that "[c]ontent that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm" are prohibited and will be removed from YouTube.[265] YouTube has further stated that it will remove content that attempts to mislead people about the voting process or any other false information relating to elections.[266] YouTube also recently created an Intelligence Desk to help review technically-manipulated content and take proactive approaches to mitigate the spread of such content,[267] and the company has also changed its recommendations systems to prevent people from viewing misinformation on its site.[268]

In the absence of formal regulation, the platforms have been left to decide for themselves where and how to draw the line between protected free speech and unprotected harmful misinformation on their platforms—and they have reached different conclusions on where that lines falls.

Where Twitter characterizes political ads as "paying for reach"[269] and does not allow them on its platform and further does not allow even non-political, cause-based ads to be

[261] Lesley Stahl, *How Does YouTube Handle the Site's Misinformation, Conspiracy Theories, and Hate?*, CBS NEWS: 60 MINUTES (Dec. 1, 2019), https://www.cbsnews.com/news/is-youtube-doing-enough-to-fight-hate-speech-and-conspiracy-theories-60-minutes-2019-12-01/.
[262] *Id.*
[263] *Id.*
[264] *How YouTube Supports Elections*, YOUTUBE OFFICIAL BLOG (Feb. 3, 2020), https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html.
[265] *Id.*
[266] *Id.*
[267] *Id.*
[268] *Id.*
[269] Jack Dorsey (@jack), *supra* note 128.

microtargeted, Facebook exempts politicians from fact-checking entirely and permits microtargeting for political ads, while allowing for the flagging and fact-checking of potential political misinformation made available by non-politicians on its platform. Facebook contends that this is the proper line to draw because political speech is subject to sufficient scrutiny among the polity and the free press,[270] notwithstanding the fact that microtargeting of ads allows politicians to avoid this broad scrutiny. In response to studies about entrenchment of false beliefs, Facebook changed its terminology on false content alerts from "disputed" to "additional reporting on this," which suggests some measure of responsiveness on Facebook's part to data about the negative impacts of the platform's policies.

Google, for its part, permits ad targeting, but only based on a limited set of characteristics, as discussed above, and does not permit some of its most powerful tools to be used for promoting political ads. Google's policies apply to YouTube and its display ad network, not merely to the eponymous search engine itself. In addition, Google prohibits ads that undermine trust in elections, as well as deepfakes or other doctored media related to "politics, social issues, or matters of public concern,"[271] which it distinguishes from mere spoken falsehoods.[272] Rather than demonetizing content that attempts to mislead people about elections or the voting process, YouTube removes the content outright.[273]

III. Analysis and Assessment of Platforms' Measures to Combat Medical and Political Misinformation

The efforts undertaken by the major social media platforms' measures to address medical and political misinformation are not without their problems. These efforts, however, are generally consistent with First Amendment substantive and procedural values, are trending in the right direction, and are by and large welcomed by the American public. The platforms' efforts are not subject to First

[270] *Fact-Checking on Facebook: Program Policies*, *supra* note 154.

[271] *Misrepresentation*, *supra* note 252.

[272] Stahl, *supra* note 261.

[273] *How YouTube Supports Elections*, *supra* note 264.

Amendment scrutiny, since the platforms are not state actors.[274] On the contrary, the platforms enjoy great discretion with respect to the choices they make regarding content regulation on their platforms, thanks to Section 230 of the Communications Decency Act (at least for now).[275] That said, the measures that the platforms have undertaken to combat misinformation have been largely consistent with First Amendment substantive and procedural values.

First, the platforms' most interventionist efforts with respect to false medical misinformation and false/misleading statements of fact in the health and medical context are consistent with First Amendment substantive values, in which lesser protection is accorded for false and misleading statements of fact (especially in the medical field).[276] While the marketplace

[274] *See*, *e.g.*, DAWN C. NUNZIATO, VIRTUAL FREEDOM: NET NEUTRALITY AND FREE SPEECH IN THE INTERNET AGE (2009).

[275] 47 U.S.C. § 230(c) (2018). The Communications Decency Act Section 230 prohibits any attempt to hold social media platforms liable for hosting harmful speech or for taking steps to remove harmful speech. *Id.* Section 230(c)(1) of the Act provides that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." *Id.* Courts have consistently interpreted this provision to immunize social media platforms from liability for hosting a variety of categories of harmful speech, including causes of action such as defamation, negligence, gross negligence, nuisance, sending threatening messages, and even statutory violations of the Fair Housing Act and related anti-discrimination violations. *See generally* Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401 (2017). In addition, the "good Samaritan" provision of Section 230 immunizes platforms from liability for undertaking measures to screen or block content on their platforms, providing that platforms cannot "be held liable on account of . . . any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene . . . excessively violent, harassing, or otherwise objectionable . . . ." 47 U.S.C. § 230(c)(2)(A). President Trump has recently taken aim at Section 230. *See* Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (June 2, 2020).; Justin Wise, *Trump to Order Review of Law Protecting Social Media Firms After Twitter Spat: Report,* THE HILL, (May 28, 2020), https://thehill.com/policy/technology/499871-trump-to-order-review-of-law-protecting-social-media-from-responsibility; Mike Masnick, *House Government Appropriations Bill Would Bar FTC & FCC From Doing Anything Related to Trump's Insane Anti-230 Executive Order,* TECH DIRT, (July 15, 2020), https://www.techdirt.com/articles/20200714/23061044903/house-government-appropriations-bill-would-bar-ftc-fcc-doing-anything-related-to-trumps-inane-anti-230-executive-order.shtml.

[276] *See* Gertz v. Robert Welch, Inc., 418 U.S. 323, 340 (1974) ("[T]here is no constitutional value in false statements of fact. Neither the intentional lie nor the careless error materially advances society's interest in 'uninhibited, robust, and wide-open' debate on public issues."); Warner-Lambert Co. v. F.T.C., 562 F.2d 749 (D.C. Cir. 1977) *enforced sub nom.* In the Matter of Warner-Lambert Co., 92 F.T.C. 191 (1978) (enjoining Listerine mouthwash advertisements unless they contained corrective language, as remedy for ads misrepresenting the efficacy of Listerine in

of ideas theory (and its default response of counter-speech as a remedy for bad speech) accords broad protection to good and bad *ideas*, it does not accord the same broad protections to good and bad claims or assertions of *fact*.[277] 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 that false statements of *fact* are not similarly immune from regulation.[278] While the Court has sometimes recognized the minimal potential contributions to the marketplace of ideas made by harmless lies,[279] or some false statements of fact,[280] it has also emphasized that the First Amendment does not stand in the way of regulating intentionally false or misleading assertions of fact,[281] especially in the medical context. Indeed, in the context of false and misleading statements of fact regarding medical treatments, cures, medicine, etc., the Food and Drug Administration (FDA) and the Federal Trade Commission (FTC) have extensive authority, consistent with the First Amendment, to prohibit false and misleading claims. The FDA and the FTC are empowered to prohibit the false or misleading branding, advertising, marketing, and/or sale of products—including products that claim to be cures or treatments for COVID-19—and these agencies have recently cracked down on

preventing, treating, or alleviating the common cold); *see generally* Nunziato, *The Marketplace of Ideas Online*, *supra* note 81.

[277] Nunziato, *The Marketplace of Ideas Online*, *supra* note 81, at 1526.

[278] *See* Gertz, 418 U.S. at 340; Nunziato, *The Marketplace of Ideas Online*, *supra* note 81, at 1526.

[279] *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. *Id.* at 724 (plurality opinion). 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). *Id.* at 715. 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. *Id.* 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. *Id.* at 719. Justice Breyer concurring in judgment, argued that Alvarez's lie was seemingly harmless and could be remedied by counterspeech. *See id.* at 732. For example, Alvarez's lie could be easily outed by a publicly accessible, online list of Medal of Honor recipients.

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

[281] *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)).

online purveyors of such products.[282] Thus, it is not inconsistent with First Amendment values for the social media platforms to undertake measures to combat false and misleading statements of fact, especially in the area of medical and health related information.

In addition, the platforms' efforts to remove content likely to incite violence or great public harm is consistent with the emergency exception in First Amendment jurisprudence, as originally articulated by Holmes and Brandeis[283] and as recognized by the Court in its incitement jurisprudence in *Brandenburg v. Ohio*[284] and its progeny. Content that is created or shared with the purpose of immediately contributing to or exacerbating violence or physical harm is generally subject to regulation under the First Amendment's incitement jurisprudence, under which the government is permitted to regulate "advocacy of . . . law violation . . . where such advocacy

[282] *See* Alexandra Sternlicht, *The FTC Has Sent Cease-And-Desist Letters to Over 150 Companies Who Claim to Have COVID-19 Cures*, FORBES (July 9, 2020), https://www.forbes.com/sites/alexandrasternlicht/2020/07/09/the-ftc-has-sent-cease-and-desist-letters-to-over-150-companies-who-claim-to-have-COVID-19-cures/#34ef5282722e (FDA has sent warning letters to over 150 companies who claim to have COVID-19 cures); Meagan Flynn, *Leader of Fake Church Peddling Bleach as COVID-19 Cure Sought Trump's Support. Instead, He Got Federal Charges.*, WASH. POST (July 9, 2020), https://www.washingtonpost.com/nation/2020/07/09/fake-coronavirus-cure-bleach/ (Criminal charges for conspiracy to defraud the United States and deliver misbranded drugs were brought against fake Florida church that claims to have COVID-19 cures). The FDA has authority to regulate purveyors of such products on the grounds that they "misleadingly" represent that their products are safe and effective for the treatment or prevention of COVID-19 and that the products are therefore illegal unapproved and misbranded products under Section 502 of the Food Drug & Cosmetic Act (the "FD&C Act"). 21 U.S.C. § 352. In addition, under the Federal Trade Commission Act (the "FTC Act"), 15 U.S.C. § 41 et seq., "it is unlawful . . . to advertise that a product can prevent, treat, or cure human disease unless the purveyor possesses competent and reliable scientific evidence, including, when appropriate, well-controlled human clinical studies, substantiating that the claims are true at the time they are made." Asahi Shimbun, *20 More Warning Letters Tell Companies to Cut Out Unproven COVID-19 Claims*, FED. TRADE COMM'N BUS. BLOG (Aug. 14, 2020, 11:41 AM), https://www.ftc.gov/news-events/blogs/business-blog/2020/08/20-more-warning-letters-tell-companies-cut-out-unproven. Accordingly, to make or exaggerate such claims without scientific evidence sufficient to substantiate them violates the FTC Act. *Id.*

[283] Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting). As Holmes explained in his *Abrams* dissent, "[o]nly the emergency that makes it immediately dangerous to leave the correction of evil counsels to time warrants making an exception to the sweeping command, 'Congress shall make no law . . . abridging the freedom of speech.'" *Id.* at 630–31; *see also* Whitney v. California, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring) ("Only an emergency can justify repression. Such must be the rule if authority is to be reconciled with freedom.").

[284] 395 U.S. 444 (1969).

is directed to inciting or producing imminent lawless action and is likely to incite or produce such action."[285]

Further, the platforms' efforts to label less harmful false and misleading medical information, and to develop and refer users to accurate information, revolves primarily around providing *counterspeech* instead of implementing censorship as a remedy. This is consistent with First Amendment substantive values and with the marketplace of ideas theory of the First Amendment, according to which—ever since the formative years of modern First Amendment jurisprudence—the accepted response to bad speech is not censorship but more (better) speech.[286] As Justice Brandeis explained in his oft-quoted concurrence in *Whitney v. California*,[287] joined by Justice Holmes: "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."[288]

According to the marketplace theory of the First Amendment, ideas should generally be allowed to compete freely in the marketplace unfettered by government restrictions (absent emergency conditions).[289] The default remedy for harmful ideas in the marketplace of speech is not censorship, but counterspeech, which operates by allowing those who are exposed to bad speech to be exposed to good speech as a counterweight.[290] The platforms' efforts to respond to false and misleading medical and political information by labeling them as such, and to refer users to accurate information, is consistent with this counterspeech approach in First Amendment jurisprudence. In addition, the platforms' efforts in regulating misinformation in political speech and political advertising contribute toward "producing an informed public capable of conducting its own affairs" and facilitating the preconditions necessary for citizens to engage in the task of democratic self-government,[291] which are also foundational goals of our First Amendment jurisprudence.

[285] *Id.* at 447.
[286] *Abrams*, 250 U.S. at 630–31 (Holmes, J., dissenting).
[287] 274 U.S. 357 (1927).
[288] *Id.* at 375–77 (1927) (Brandeis, J., concurring) (emphasis added).
[289] Nunziato, *The Marketplace of Ideas Online*, *supra* note 81, at 1520–21.
[290] *Id.*
[291] Red Lion Broad. Co. v. F.C.C., 395 U.S. 367, 392 (1960).

The platforms' efforts are also generally consistent with First Amendment *procedural* values and with principles of due process generally.[292] The First Amendment's protections for freedom of expression not only embody a substantive dimension of which categories of speech to protect; they also embody procedural dimensions, imported from the Due Process Clause, which require that "sensitive tools" be implemented by decisionmakers in restricting speech.[293] As free speech theorist Henry Monaghan explains, "procedural guarantees play an equally large role in protecting freedom of speech; indeed, they assume an importance fully as great as the validity of the substantive rule of law to be applied."[294]

Accordingly, First Amendment jurisprudence incorporates a powerful "body of procedural law which defines the manner in which [decisionmakers] must evaluate and resolve [free speech] claims —[establishing] a First Amendment due process."[295] This jurisprudence embodies "a comprehensive system of procedural safeguards designed to obviate the dangers of a censorship system."[296] Consistent with these procedural safeguards embodied in First Amendment jurisprudence, social media platforms should impose speech restrictions on medical and political misinformation in a clear, neutral, and transparent manner such that speakers are adequately and clearly informed of the platforms' rules regarding speech, speakers are specifically informed of the reasons why their speech was restricted (removed or labeled), decisions are made consistently by impartial decisionmakers, and speakers have an opportunity to be heard to appeal any such speech restrictions. In general, the platforms have provided clear notice to users of their (evolving) terms of service regarding medical and political misinformation and have provided users with clear notice when implementing speech removal or labeling decisions. For example, as discussed above, when Twitter restricted Donald Trump, Jr.'s posts embodying false claims about unproven cures for COVID-19 on

[292] *See generally*, Dawn Carla Nunziato, *How (Not) To Censor: Procedural First Amendment Values and Internet Censorship Worldwide*, 42 GEO. J. INT'L L. 1123 (2014); Dawn Carla Nunziato, *Forget About It? Harmonizing European and American Protections for Privacy, Free Speech, and Due Process*, *in* PRIVACY AND POWER (Cambridge University Press, 2017).

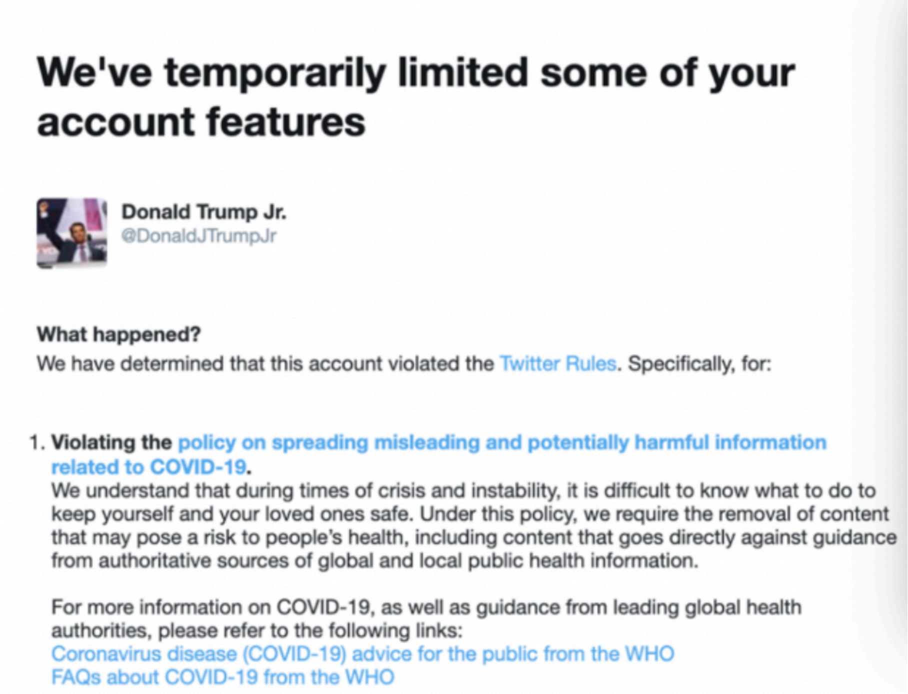
[293] Bantam Books v. Sullivan, 372 U.S. 58, 66 (1963).

[294] Henry Monaghan, *First Amendment "Due Process"*, 83 HARV. L. REV. 518, 518 (1970) (internal quotations omitted).

[295] *Id.*

[296] *Id.* (internal quotations omitted).

the grounds that the post violated Twitter's rules regarding medical misinformation,[297] it did so in the context of providing clear prior notice of what speech was restricted and a process to appeal Twitter's decisions,[298] and it also provided notice to Trump, Jr., of the specific reason why his speech was restricted. See below.



We've temporarily limited some of your account features

Donald Trump Jr.
@DonaldJTrumpJr

What happened?
We have determined that this account violated the Twitter Rules. Specifically, for:

1. **Violating the policy on spreading misleading and potentially harmful information related to COVID-19.**
We understand that during times of crisis and instability, it is difficult to know what to do to keep yourself and your loved ones safe. Under this policy, we require the removal of content that may pose a risk to people's health, including content that goes directly against guidance from authoritative sources of global and local public health information.

For more information on COVID-19, as well as guidance from leading global health authorities, please refer to the following links:
Coronavirus disease (COVID-19) advice for the public from the WHO
FAQs about COVID-19 from the WHO

Figure Thirteen: Twitter's notice to Donald Trump Jr. after restricting Trump Jr.'s account features.[299]

In short, the extensive measures undertaken by the major social media platforms to respond to false and misleading misinformation in the medical and political contexts are generally consistent with our First Amendment substantive and procedural values.

[297] *See supra* notes 54–58 and accompanying text.

[298] *See Appeal an Account Suspension or Locked Account*, TWITTER, https://help.twitter.com/forms/general?subtopic=suspended (last accessed Sept. 5, 2020) (setting forth the procedural for users to appeal severe violations of Twitter's rules resulting in suspending and/or blocked accounts).

[299] Katelyn Caralle, *Twitter stops Donald Trump Jr. from tweeting for 12 hours after he promoted doctor's claim hydroxychloroquine 'cures' COVID and called it 'must-watch' - on eve of big tech bosses being quizzed by Congress,* DAILYMAIL.COM (July 28, 2020), https://www.dailymail.co.uk/news/article-8568579/Twitter-cancels-Don-Jr-s-account-access-posting-claim-hydroxychloroquine-cures-COVID.html.

In addition, recent studies have shown that the efforts undertaken by the major social media platforms' measures to address political and medical misinformation have been moderately successful. As 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."[300] The authors of the study find that "user interaction with known false news sites has declined by 50[%] since the 2016 election."[301] 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'[302]] may have had a meaningful impact."[303]

Further, the labeling of content as false or misleading on social media platforms has been shown to be effective in limiting the dissemination of false or misleading content. According to a recent study, social media users were about 50% less likely to share false stories if the stories had been labeled as false.[304] When no labels were used at all, participants considered sharing 29.8% of false stories in the sample, but that figure dropped to 16.1% of false stories that had a label attached.[305] In addition, 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 60% of respondents gave accurate answers when presented with a fact-check/correction, while only 32% expressed accurate

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

[301] *Id.* at 5.

[302] *Id.* at 6.

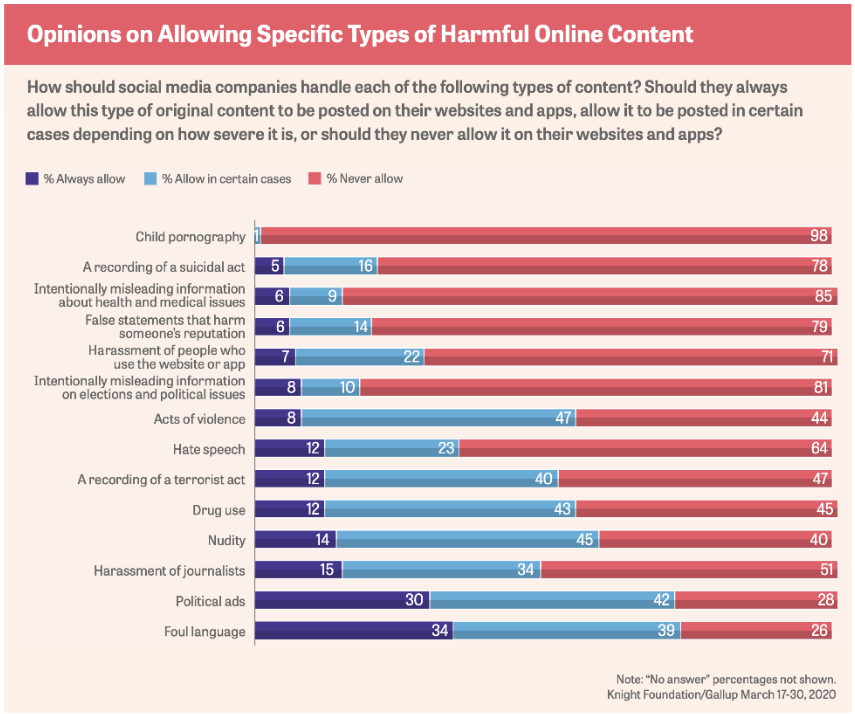
[303] *Id.* at 3.

[304] *See* 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.

[305] *Id.*

beliefs when they were not presented with a fact-check/correction.[306]

Finally, there is broad public support among Americans for social media platforms' continuing to take a meaningful role in combating political and medical misinformation on their platforms. A March 2020 Knight Foundation/Gallup Poll found that the vast majority of Americans surveyed (81%) supported the removal of intentionally misleading information on elections or other political issues, and an even greater majority of Americans surveyed (85%) supported social media companies' removal of intentionally misleading health information.[307]



Opinions on Allowing Specific Types of Harmful Online Content

How should social media companies handle each of the following types of content? Should they always allow this type of original content to be posted on their websites and apps, allow it to be posted in certain cases depending on how severe it is, or should they never allow it on their websites and apps?

Note: "No answer" percentages not shown.
Knight Foundation/Gallup March 17-30, 2020

Figure Fourteen: March 2020 data from a Knight Foundation/Gallup Poll measuring public opinions about harmful content online.[308]

[306] Drutman, *supra* note 12. 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.*

[307] FREE EXPRESSION, HARMFUL SPEECH AND CENSORSHIP IN A DIGITAL WORLD, supra note 11, AT 6.

[308] *Id.* at 6.

CONCLUSION

Social media platforms are playing an ever-expanding role in shaping the contours of the information ecosystem today, as these platforms have shouldered the burden of ensuring that the public is informed—and not misinformed—about matters affecting our democratic institutions in the context of our elections, as well as about matters affecting our very health and lives, in the context of the pandemic. The platforms are attempting to address these serious problems alone, in the absence of federal or state regulation or guidance in the United States. While the platforms' intervention in the online marketplace of ideas is not without its problems, this Article has argued that this intervention is by and large a salutary development and is one that has brought about improvements in the online information ecosystem. Social media companies have been generally inspired by First Amendment free speech values––both substantive and procedural—to protect a vibrant marketplace of ideas online while imposing limited, moderately effective checks on harmful false and misleading speech, with complex systems of removal, fact-checking, and labeling, and by serving up prominent information from independent fact-checkers and trusted authorities to counter medical and political misinformation. In the absence of effective regulatory measures in the United States to combat medical and political misinformation online, social media companies should be commended for their efforts thus far and should continue to develop and deploy even more successful measures to combat such misinformation online.