MODERATING CONTENT MODERATION: A FRAMEWORK FOR NONPARTISANSHIP IN ONLINE GOVERNANCE

EDWARD LEE*

Internet platforms serve two important roles that often conflict. Facebook, Twitter, YouTube, and other internet platforms facilitate the unfettered exchange of free speech by millions of people, yet they also moderate or restrict the speech according to their "community standards," such as prohibitions against hate speech and advocating violence, to provide a safe environment for their users. These dual roles give internet platforms unparalleled power over online speech—even more so than most governments. Yet, unlike government actors, internet platforms are not subject to checks and balances that courts or agencies must follow, such as promulgating well-defined procedural rules and affording notice, due process, and appellate review to individuals. Internet platforms have devised their own policies and procedures for content moderation, but the platforms' enforcement remains opaque—especially when compared to courts and agencies. Based on an independent survey of the community standards of the largest internet platforms, this Article shows that few internet platforms disclose the precise procedural steps and safeguards of their content moderation—perhaps hoping to avoid public scrutiny over those procedures. This lack of transparency has left internet platforms vulnerable to vocal accusations of having an "anti-conservative bias" in their content moderation, especially from politicians. Internet platforms deny such a bias, but their response has not mollified Republican lawmakers, who have proposed amending, if not repealing, Section 230 of the Communications Decency Act

* Professor of Law, *IIT Chicago-Kent College of Law*; Founder, The Free Internet Project. Many thanks to helpful comments from Kathy Baker, Felice Batlan, Sungjoon Cho, Eric Goldman, Hal Krent, Nancy Marder, Blake Reid, Mark Rosen, Alex Boni-Saenz, Chris Schmidt, Stephanie Stern, Eugene Volokh, and participants of faculty workshops. This Article represents my own views as a legal scholar. They should not be attributed to the nonprofit The Free Internet Project.

to limit the permissible bases and scope of content moderation that qualify for civil immunity under the section. This Article provides a better solution to this perceived problem—a model framework for nonpartisan content moderation (NCM) that internet platforms should voluntarily adopt as a matter of best practices. The NCM framework provides greater transparency and safeguards to ensure nonpartisan content moderation in a way that avoids messy government entanglement in enforcing speech codes online. The NCM framework is an innovative approach to online governance that draws upon safeguards designed to promote impartiality in various sectors, including courts and agencies, clinical trials, peer review, and equal protection under the Fourteenth Amendment.

TABLE OF CONTENTS

No man is allowed to be a judge in his own cause; because his interest would certainly bias his judgment, and, not improbably, corrupt his integrity.
—Madison, FEDERALIST NO. 10

Were there not even these inducements to moderation, nothing can be more ill-judged than that intolerant spirit, which has, at all times, characterized political parties.
—Hamilton, FEDERALIST NO. 1

INTRODUCTION

In 2020, amidst a pandemic and nationwide protests led by Black Lives Matter following the brutal police killing of George Floyd, internet platforms[1] tightened their policies of content moderation—otherwise known as "community standards"—to stop the spread of

1. The term "internet platform" is an evolving, even "slippery term." TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET 18 (2018). I borrow Tarleton Gillespie's definition: "online sites and services that (a) host, organize, and circulate users' shared content or social interactions for them, (b) without having produced or commissioned (the bulk of) that content, (c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit." *Id.*

misinformation, hate speech, and voter suppression.[2] The platforms had implemented new policies to curb foreign interference and misinformation that were pervasive in the 2016 U.S. election,[3] but the platforms took a hands-off approach to the content of U.S. politicians and political ads. That changed in 2020. On May 26, 2020, as protests over Floyd's death erupted, Twitter started the sea change by flagging several of President Donald Trump's tweets with labels indicating that his tweets violated Twitter's policies against misinformation, voter suppression, and glorification of violence.[4] Snapchat and Twitch followed suit by announcing their own efforts to moderate or outright suspend Trump's accounts on their respective platforms due to concerns about "amplify[ing] voices who incite racial violence and injustice"[5] and "hateful conduct."[6] Reddit, known for its über-permissiveness, even banned a subreddit, or discussion group, devoted to "r/The Donald"

---

2. *See* Craig Timberg & Elizabeth Dwoskin, *Silicon Valley Is Getting Tougher on Trump and His Supporters over Hate Speech and Disinformation*, WASH. POST (July 10, 2020, 1:53 PM), https://www.washingtonpost.com/technology/2020/07/10/hate-speech-trump-tech; Barbara Ortutay & Tali Arbel, *Social Media Platforms Face a Reckoning over Hate Speech*, AP NEWS (June 29, 2020, 6:00 PM), https://apnews.com/article/6d0b3359ee5379bd5624c9f1024a0eaf.

3. *See infra* Section I.B.1.

4. *See Trump Makes Unsubstantiated Claim that Mail-in Ballots Will Lead to Voter Fraud*, TWITTER (May 26, 2020), https://twitter.com/i/events/1265330601034256384?lang=en; Barbara Sprunt, *The History Behind 'When the Looting Starts, the Shooting Starts*,' NPR (May 29, 2020, 6:45 PM), https://www.npr.org/2020/05/29/864818368/the-history-behind-when-the-looting-starts-the-shooting-starts [https://perma.cc/N4MZ-3E82]; William Mansell & Libby Cathey, *Twitter Flags Trump, White House for 'Glorifying Violence' in Tweets About George Floyd Protests*, ABC NEWS (May 29, 2020, 2:32 PM), https://abcnews.go.com/US/twitter-flags-trump-white-house-glorifying-violence-tweet/story?id=70945228 [https://perma.cc/6Y3C-BYJL]; *Twitter Flags Trump's Tweet of Doctored 'Racist Baby' Video*, AP NEWS (June 19, 2020), https://apnews.com/3499484ab404647b01fcc4a08babff03; Lauren Feiner, *Twitter Flagged Another Trump Tweet for Violating Its Policies*, CNBC (June 23, 2020, 6:27 PM), https://www.cnbc.com/2020/06/23/twitter-labeled-another-trump-tweet-for-violating-its-policies.html [https://perma.cc/VPV8-4KKH]. Some of Trump's tweets were removed due to claims of copyright infringement. *See Twitter Disables Trump Tweet over Copyright Complaint*, REUTERS (July 18, 2020, 11:06 PM), https://www.reuters.com/article/us-usa-trump-twitter/twitter-disables-trump-tweet-over-copyright-complaint-idUSKCN24K02U.

5. *See* Salvador Rodriguez, *Snap Will No Longer Promote Trump Posts Within Snapchat*, CNBC (June 3, 2020, 1:48 PM), https://www.cnbc.com/2020/06/03/snapchat-will-no-longer-promote-trump-posts.html [https://perma.cc/RNX3-7QEM].

6. Kellen Browning, *Twitch Suspends Trump's Channel for 'Hateful Conduct,'* N.Y. TIMES (June 30, 2020), https://www.nytimes.com/2020/06/29/technology/twitch-trump.html.

for violating Reddit's new policy against "promoting hate based on identity or vulnerability."[7] Facebook stepped up its moderation and revised its lax policy after more than 400 companies waged a nationwide ad boycott "Stop Hate for Profit."[8] CEO Mark Zuckerberg said Facebook will add labels to content—including from politicians—that violates its community standards but is allowed to remain on Facebook under a "newsworthiness" exception.[9] Facebook removed a Trump campaign ad that contained a symbol associated with a Nazi symbol.[10] In response, Trump and prominent Republicans decried the "censorship," "election interference," and "anti-conservative bias" of internet platforms.[11] Republican lawmakers also promoted a social media company named Parler, which touted itself as "unbiased" social media, even though it reportedly began removing users with liberal views.[12]

The controversy intensified during and after the 2020 U.S. election. Twitter, Facebook, and other platforms implemented new measures to stop election misinformation, especially false claims of victory and false

---

7. *Update to Our Content Policy*, REDDIT (June 29, 2020), https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy [https://perma.cc/QRQ8-KK54]; Mike Isaac, *Reddit, Acting Against Hate Speech, Bans 'The_Donald' Subreddit*, N.Y. TIMES (June 30, 2020), https://www.nytimes.com/2020/06/29/technology/reddit-hate-speech.html.

8. Elizabeth Culliford & Sheila Dang, *Facebook Will Label Newsworthy Posts that Break Rules as Ad Boycott Widens*, REUTERS (June 26, 2020, 12:59 PM), https://www.reuters.com/article/us-facebook-ads-boycott-unilever/facebook-will-label-newsworthy-posts-that-break-rules-as-ad-boycott-widens-idUSKBN23X2FW.

9. *See* Mark Zuckerberg, FACEBOOK (June 26, 2020, 11:25 AM), https://www.facebook.com/zuck/posts/10112048980882521 [https://perma.cc/CBX7-P4ZD]; Rachel Sandler, *In Reversal, Zuckerberg Says Facebook Will Label Newsworthy Posts that Violate Its Rules*, FORBES (June 26, 2020, 5:11 PM), https://www.forbes.com/sites/rachelsandler/2020/06/26/in-reversal-zuckerberg-says-facebook-will-label-newsworthy-posts-that-violate-its-rules/#4bc711407340 [https://perma.cc/3MQ3-CKLJ].

10. *See* Annie Karni, *Facebook Removes Trump Ads Displaying Symbol Used by Nazis*, N.Y. TIMES (June 30, 2020), https://www.nytimes.com/2020/06/18/us/politics/facebook-trump-ads-antifa-red-triangle.html.

11. *See* Brian Fung et al., *Trump Signs Executive Order Targeting Social Media Companies*, CNN (May 28, 2020, 9:22 PM), https://www.cnn.com/2020/05/28/politics/trump-twitter-social-media-executive-order/index.html [https://perma.cc/SQ6A-YRD8].

12. *See* Marina Watts, *Parler, the Ted Cruz-Approved 'Free Speech' App, Is Already Banning Users*, NEWSWEEK (June 30, 2020, 11:22 AM), https://www.newsweek.com/parler-ted-cruz-approved-free-speech-app-already-banning-users-1514358.

claims of voter fraud, including from the candidates.[13] By two days after the election, Twitter had added warning labels to thirty-eight percent of Trump's tweets that claimed both voter fraud and victories for him in several states that had yet to be called; the labels indicated that "[s]ome or all of the content shared in this Tweet is disputed and might be misleading about an election or other civic process."[14]

The coup de grâce came on January 6, 2021, when a group of violent insurrectionists spurred on by Trump[15] broke into the U.S. Capitol in an unsuccessful attempt to stop Congress from certifying the election of Joseph Biden as the next President.[16] Instead of condemning the insurrection, which led to at least five deaths including a Capitol police officer,[17] Trump initially tweeted a video expressing his "love" for the "very special" people, meaning the insurrectionists who attacked the Capitol, while asking them to "go home" and again falsely claiming the election was "stolen" from him.[18] That prompted the sternest response from Facebook, Twitter,

---

13. *See* Ellen P. Goodman & Karen Kornbluh, *How Well Did Twitter, Facebook, and YouTube Handle Election Misinformation?*, SLATE (Nov. 10, 2020, 7:12 PM), https://slate.com/technology/2020/11/twitter-facebook-youtube-misinformation-election.html [https://perma.cc/V9TW-7UCL].

14. *See* Kate Conger, *Twitter Has Labeled 38% of Trump's Tweets Since Tuesday*, N.Y. TIMES (Nov. 5, 2020), https://www.nytimes.com/2020/11/05/technology/donald-trump-twitter.html.

15. *See* Rebecca Ballhaus et al., *Trump and His Allies Set the Stage for Riot Well Before January 6*, WALL ST. J. (Jan. 8, 2021, 8:38 PM), https://www.wsj.com/articles/trump-and-his-allies-set-the-stage-for-riot-well-before-january-6-11610156283; Maggie Haberman, *Trump Told Crowd 'You Will Never Take Back Our Country with Weakness,'* N.Y. TIMES (Jan. 6, 2021), https://www.nytimes.com/2021/01/06/us/politics/trump-speech-capitol.html; Dan Mangan, *Education Secretary Betsy DeVos Resigns over Capitol Riot, Blames Trump Rhetoric*, CNBC (Jan. 7, 2021, 9:05 PM), https://www.cnbc.com/2021/01/07/education-secretary-betsy-devos-resigns-over-capitol-riot-blames-trump-rhetoric.html [https://perma.cc/2LWF-9LNX]; *Trump Promises 'Wild' Protest in DC on Jan. 6, the Day Congress to Count Electoral Votes*, FOX 5 (Dec. 19, 2020), https://www.fox5dc.com/news/trump-promises-wild-protest-in-dc-on-jan-6-the-day-congress-to-count-electoral-votes [https://perma.cc/7R22-SJZC].

16. *See How Pro-Trump Insurrectionists Broke into the U.S. Capitol*, WASH. POST (Jan. 6, 2021), https://www.washingtonpost.com/politics/interactive/2021/video-timeline-capitol-breach.

17. *See* Dartunorro Clark & Frank Thorp V, *Capitol Police Officer Dies from Injuries After Clashing with Pro-Trump Mob*, NBC NEWS (Jan. 8, 2021, 7:15 AM), https://www.nbcnews.com/politics/politics-news/capitol-police-officer-has-died-after-clashing-pro-trump-mob-n1253396 [https://perma.cc/HSV8-ML42].

18. *See 'Go Home. We Love You.': Trump to Protesters*, ABC NEWS, https://abcn

and other platforms that feared further incitement of violence: the removal of Trump's video and suspension of his social media accounts.[19] Trump was forever barred from tweeting, something he did frequently during his presidency.[20]

Even before 2020, content moderation by internet platforms was contentious. In 2016, *Gizmodo* published an article in which unnamed, former Facebook employees alleged that Facebook "prevented stories about the right-wing CPAC gathering . . . and other conservative topics from appearing in the highly-influential [newsfeed on Facebook], even though they were organically trending among the site's users."[21] Tom Stocky, Facebook's Vice President of Search, issued a categorical denial: "There are rigorous guidelines in place for the review team to ensure consistency and neutrality. These guidelines do not permit the suppression of political perspectives."[22] Zuckerberg reiterated the company's policy to incredulous Republican lawmakers at contentious hearings in April 2018[23] and April 2019.[24] Although Facebook's alleged censorship was not the

---

.ws/3nkLqSk; Salvador Rodriguez, *Facebook, Twitter Lock Trump's Account Following Video Addressing Washington Rioters*, CNBC (Jan. 7, 2021, 9:04 AM), https://www.cnbc.com/2021/01/06/twitter-pledges-action-on-any-calls-for-violence-in-capitol-riot.html [https://perma.cc/5YPB-ASV4].

19. *See* Rodriguez, *supra* note 18; James Clayton et al., *Trump Allowed back onto Twitter*, BBC (Jan. 8, 2021), https://www.bbc.com/news/technology-55569604 [https://perma.cc/J6VV-BUYX]; Sara Fischer & Ashley Gold, *All the Platforms that Have Banned or Restricted Trump so far*, AXIOS (Jan. 11, 2021), https://www.axios.com/platforms-social-media-ban-restrict-trump-d9e44f3c-8366-4ba9-a8a1-7f3114f920f1.html [https://perma.cc/RYB9-S4PK].

20. *See* Fischer & Gold, *supra* note 19; *Rate of Tweets (Donald Trump on Social Media)*, WIKIPEDIA, https://en.wikipedia.org/wiki/Donald_Trump_on_social_media#Rate_of_tweets [https://perma.cc/NR64-NFHA] (averaging 34.8 tweets per day from July 20, 2020 to January 8, 2021).

21. Michael Nunez, *Former Facebook Workers: We Routinely Suppressed Conservative News*, GIZMODO (May 9, 2016, 4:10 PM), https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006.

22. *Id.*; Tom Stocky, FACEBOOK (May 9, 2016), https://www.facebook.com/tstocky/posts/10100853082337958 [https://perma.cc/ERK6-D7KL].

23. *See* Kathleen Chaykowski, *Congressional Leaders Press Zuckerberg on Political Bias, Data Collection at Facebook*, FORBES (Apr. 11, 2018, 5:50 PM), https://www.forbes.com/sites/kathleenchaykowski/2018/04/11/congressional-leaders-press-zuckerberg-on-political-bias-data-collection-at-facebook/#5b8c3e231a95 [https://perma.cc/GLW2-LHXY]; *Facebook, Social Media Privacy, and the Use and Abuse of Data: J. Hearing Before the S. Comm. on Commerce, Science, & Transportation & on the Judiciary*, 115th Cong. (2018).

24. *See* David Shepardson, *Facebook, Google Accused of Anti-Conservative Bias at U.S. Senate Hearing*, REUTERS (Apr. 10, 2019, 5:35 PM), https://www.reuters.com/article/

focus of the 2018 hearing, within a year, anti-conservative bias went from a secondary issue to the main event at the congressional hearing tendentiously titled, "Stifling Free Speech: Technological Censorship and the Public Discourse."[25]

The internet platforms' increase in content moderation of Trump in 2020 provoked swift responses. Trump issued the Executive Order on Preventing Online Censorship[26] that interprets "good faith" in Section 230 of the Communications Decency Act[27] (CDA) to require internet services to avoid moderating user content in a way that is "deceptive, pretextual, or inconsistent with a provider's terms of service," including "deceptive or pretextual actions . . . to stifle viewpoints with which they disagree."[28] If the internet service violates this requirement, it should lose Section 230 immunity,[29] which broadly protects internet services from civil liability based on the content posted by their users or based on the companies' "good faith" screening of "objectionable" content.[30] The Department of Justice (DOJ) issued a review of Section 230 that sets forth recommendations for Congress to consider, including "adding a statutory definition of 'good faith,' which would limit immunity for content moderation decisions to those done in accordance with plain and particular terms of service and accompanied by a reasonable explanation."[31] Meanwhile, Republican lawmakers proposed no fewer than seven bills that would disqualify, by different mechanisms, internet platforms from Section 230 immunity for politically biased content moderation.[32] After losing but continuing to contest the

---

us-usa-congress-socialmedia/facebook-google-accused-of-anti-conservative-bias-at-u-s-senate-hearing-idUSKCN1RM2SJ.

25. Stifling Free Speech: Technological Censorship and the Public Discourse: Hearing Before the S. Judiciary Comm., 116th Cong. (2019).

26. Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (May 28, 2020).

27. 47 U.S.C. § 230.

28. *Executive Order on Preventing Online Censorship*, WHITE HOUSE (May 28, 2020), https://www.whitehouse.gov/presidential-actions/executive-order-preventing-online-censorship [https://perma.cc/J5YL-VSGB].

29. Exec. Order No. 13,925 § 2(a), 85 Fed. Reg. at 34,080.

30. 47 U.S.C. § 230(c).

31. U.S. Dep't of Just., Section 230—Nurturing Innovation or Fostering Unaccountability? (2020).

32. *See* Adam Wolfe, *Summary of EARN IT and Proposed Bills to Amend Section 230 of CDA Regarding ISP Safe Harbor*, FREE INTERNET PROJECT (June 27, 2020), https://thefreeinternetproject.org/blog/summary-earn-it-and-proposed-bills-amend-section-230-cda-regarding-isp-safe-harbor [https://perma.cc/73XH-UG25].

November 2020 election, Trump threatened to veto a defense spending bill if Congress did not repeal Section 230, which Trump decried, in a tweet, as "a serious threat to our National Security & Election Integrity."[33]

Facebook has also been criticized from the opposite end of the political spectrum: for allegedly showing favoritism to conservative politicians, including by Facebook executives reportedly intervening to stop the company's own content moderators from flagging violations by high-profile conservatives.[34] Top executives at Facebook, including Zuckerberg, Facebook's Vice President for Global Public Policy Joel Kaplan, and board member and investor Peter Thiel (Kaplan and Thiel are both Republicans), have been accused of steering, if not skewing, Facebook's content moderation policy to protect the content of conservatives on Facebook.[35] Facebook has a significant business interest in protecting and catering to conservatives: "Facebook's internal data showed that conservative voices are consistently the most popular on the site."[36]

Political bias wasn't the only charge that Facebook and other platforms faced in 2020. The nationwide protests following the killing of George Floyd led to a mass boycott by businesses that stopped all advertising on Facebook until Facebook agreed to engage in *more* content moderation to stop hate, misinformation, and bias, and to remove groups promoting white supremacy, Holocaust denialism,

---

33. *See* Jaclyn Diaz, *Trump Vows to Veto Defense Bill Unless Shield for Big Tech Is Scrapped*, NPR (Dec. 2, 2020, 6:25 AM), https://www.npr.org/2020/12/02/941019533/trump-vows-to-veto-defense-bill-unless-shield-for-big-tech-is-scrapped [https://perma.cc/P52N-JQRR].

34. *See, e.g.*, Sarah Frier & Kurt Wagner, *Facebook Needs Trump even More than Trump Needs Facebook*, BLOOMBERG BUSINESSWEEK (Sept. 17, 2020, 5:30 PM), https://www.bloomberg.com/news/features/2020-09-17/facebook-and-mark-zuckerberg-need-trump-even-more-than-trump-needs-facebook (explaining why Facebook may be working to appease President Trump); Olivia Solon, *Sensitive to Claims of Bias, Facebook Relaxed Misinformation Rules for Conservative Pages*, NBC NEWS (Aug. 7, 2020, 3:31 PM), https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182 [https://perma.cc/HTD2-QHMS].

35. *See* Frier & Wagner, *supra* note 34; Solon, *supra* note 34; Emily Glazer et al., *Peter Thiel at Center of Facebook's Internal Divisions on Politics*, WALL ST. J. (Dec. 17, 2019, 5:30 AM), https://www.wsj.com/articles/peter-thiel-at-center-of-facebooks-internal-divisions-on-politics-11576578601; Deepa Seetharaman, *Facebook's Lonely Conservative Takes on a Power Position*, WALL ST. J. (Dec. 23, 2018, 8:00 AM), https://www.wsj.com/articles/facebooks-lonely-conservative-takes-on-a-power-position-11545570000.

36. *See* Frier & Wagner, *supra* note 34.

vaccine misinformation, and climate denialism.[37] Several Facebook employees quit and others held a virtual walkout to protest the company's failure to address the perceived problem of hate speech on Facebook.[38] The controversy boiled over when Zuckerberg admitted that Facebook made a mistake[39] in not removing a militia group's Facebook page that publicized an event asking for "patriots willing *to take up arms* and defend (Kenosha[, Wisconsin]) *from the evil thugs*."[40] The call to arms was a response to the protests following a video showing a Kenosha police officer shoot Jacob Blake, a Black man, seven times in the back.[41] On August 25, 2020, traveling from Illinois to Kenosha, 17-year-old Kyle Rittenhouse shot and killed two protesters with an AR-15-style rifle in what his lawyer claims was self-defense; Rittenhouse has been charged with multiple counts of homicide.[42] It's unclear if Rittenhouse saw the militia group's

37. *See, e.g.*, *Recommended Next Steps*, STOP HATE FOR PROFIT, https://www.stophateforprofit.org/productrecommendations [https://perma.cc/TZY7-NCE9] (recommending steps for Facebook to take to combat the proliferation of hate speech and misinformation on the platform); Jessica Guynn, *Boycott Facebook: Civil Rights Groups Call on Big Advertisers to Yank Ads over Hate Speech Policies*, USA TODAY (June 18, 2020, 9:25 AM), https://www.usatoday.com/story/tech/2020/06/17/facebook-hate-speech-civil-rights-groups-call-advertising-boycott/3207915001 [https://perma.cc/Z4V7-UK35].

38. *See* Alison Durkee, *Facbeook Engineer Resigns, Says Company on 'Wrong Side of History' as Internal Dissent Grows*, FORBES (Sept. 8, 2020, 3:13 PM), https://www.forbes.com/sites/alisondurkee/2020/09/08/facebook-engineer-resigns-company-on-wrong-side-of-history-internal-employee-dissent-grows/#7c74c97d3794 [https://perma.cc/67R6-RF53].

39. *See* Todd Spangler, *Mark Zuckerberg Admits Facebook 'Operational Mistake' in Failing to Pull Wisconsin Militia Group*, VARIETY (Aug. 29, 2020, 6:36 AM), https://variety.com/2020/digital/news/mark-zuckerberg-facebook-operational-mistake-wisconsin-militia-group-1234753545.

40. Katherine Rosenberg-Douglas, *Fledgling Militia Group Put out Call to Arms in Kenosha and 5,000 People Responded. Now It's Banned from Facebook After Fatal Shootings During Protests*, CHI. TRIB. (Aug. 28, 2020, 6:40 AM) (emphasis added), https://www.chicagotribune.com/news/breaking/ct-kenosha-wisconsin-militia-social-media-shooting-20200828-aenx5ropmrfmtca34ezqvhwe7e-story.html.

41. *Id.*; *Jacob Blake: What We Know About Wisconsin Police Shooting*, BBC NEWS (Aug. 31, 2020), https://www.bbc.com/news/world-us-canada-53909766 [https://perma.cc/MN8U-MRFT]. Prosecutors decided not to press charges against the police officer Rusten Sheskey. *See* Robert Chiarito et al., *Jacob Blake Shooting: No Charges Against Officer in Kenosha, Wisconsin*, N.Y. TIMES (Jan. 5, 2021), https://nyti.ms/2Xdq52H.

42. *See* Christina Maxouris et al., *Kenosha Shooting Suspect Faces More Homicide Charges*, CNN (Aug. 27, 2020, 11:26 PM), https://www.cnn.com/2020/08/27/us/kenosha-wisconsin-shooting-suspect/index.html [https://perma.cc/GHT7-49TG]; Phil Helsel, *Kyle Rittenhouse, Charged with Killing 2 Kenosha Protesters, Extradited to*

Facebook page, but many did: "1,000 people responded they were 'going' to the Kenosha event. Another 4,000 said they were 'interested.'"[43] In addition, in what might be perceived as an effort to undermine the protesters, the Department of Homeland Security called upon internet platforms to take "appropriate action . . . against content that promotes, incites, or assists the commission of imminent illegal activities and violence," including "calls to break city curfews, information about which stores or neighborhoods to target for looting or destruction, and coordination of attacks against particular people or groups of people."[44]

It's easy to see how different goals of content moderation may result in decisions that appear "biased" depending on the viewer. Representative Steve King, a Republican from Iowa, tweeted on December 8, 2017: "Diversity is not our strength. Hungarian Prime Minister Victor Orban, 'Mixing cultures will not lead to a higher quality of life but a lower one.'"[45] King has been criticized as supporting white supremacy, even by members of his own party.[46] Yet Twitter didn't moderate King's tweet. Twitter's inaction can be criticized as helping to promote white supremacy on Twitter.[47] However, had Twitter removed the content for violating its community standards, Twitter might have been viewed as biased against conservative politicians. Or consider when Twitter moderated Trump's controversial tweet echoing former police chief Walter Headley's infamous 1967 line:

Wisconsin, U.S. NEWS (Oct. 30, 2020, 10:36 PM), https://www.nbcnews.com/news/us-news/kyle-rittenhouse-charged-killing-2-kenosha-protesters-extradited-wisconsin-n1245579 [https://perma.cc/GW6N-59B5].

43. Rosenberg-Douglas, *supra* note 40.

44. Lauren Feiner, *DHS Asks Facebook, Twitter and Others to Take Action on Posts Calling for 'Violence' amid Nationwide Protests*, CNBC (June 26, 2020, 6:24 PM), https://www.cnbc.com/2020/06/26/dhs-to-facebook-twitter-take-action-on-posts-calling-for-violence.html [https://perma.cc/3Q9K-LSLU].

45. Steve King (@SteveKingIA), TWITTER (Dec. 8, 2017, 8:00 AM), https://twitter.com/SteveKingIA/status/939117527375990790.

46. *See* Trip Gabriel, *A Timeline of Steve King's Racist Remarks and Divisive Actions*, N.Y. TIMES (Jan. 15, 2019), https://www.nytimes.com/2019/01/15/us/politics/steve-king-offensive-quotes.html.

47. *See* Grace Panetta, *Twitter Reportedly Won't Use an Algorithm to Crack down on White Supremacists Because Some GOP Politicians Could End up Getting Barred too*, BUS. INSIDER (Apr. 25, 2019, 2:19 PM), https://www.businessinsider.com/twitter-algorithm-crackdown-white-supremacy-gop-politicians-report-2019-4 [https://perma.cc/W726-XK2N].

> 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![48]

Twitter let Trump's tweet remain on its platform under what it calls (without irony) its "public[] interest" exception,[49] but flagged the tweet with a label that it "violated the Twitter Rules about glorifying violence."[50] Trump decried Twitter's moderation as an "unchecked power to censor": "Twitter ceases to be a neutral public platform."[51] But, had Twitter done nothing, the inaction would have prompted criticism from those who interpret "when the looting starts, the shooting starts" as a racist threat used in the past to commit violence against the Black community.[52] In short, internet platforms are often whipsawed: damned if they do moderate, and damned if they don't.

As is apparent from this firestorm, content moderation is the third rail of internet policy. Staking out any position is likely to elicit criticism, if not condemnation or worse. Debating content moderation has become as charged as the polarized discussions on social media. Yet the internet platforms have done little to defuse the situation. Even though Facebook, Google, and Twitter have denied having a political bias, their community standards—surprisingly—say next to nothing about the goal of nonpartisanship in content moderation of public officials and political candidates, let alone the company's procedures to ensure nonpartisanship.[53]

---

48. *See* Tommy Beer, *Trump Defends Controversial 'Shooting' Tweet, and White House Claims Twitter Admits Mistake*, FORBES (May 29, 2020, 6:59 PM), https://www.forbes.com/sites/tommybeer/2020/05/29/trump-defends-controversial-tweet-sanctioned-by-twitter/#652237d541d7 [https://perma.cc/3ENT-GFAT].

49. *About Public-Interest Exceptions on Twitter*, TWITTER, https://help.twitter.com/en/rules-and-policies/public-interest [https://perma.cc/78K2-4FAL].

50. *See* Donald J. Trump (@realDonaldTrump), TWITTER (May 29, 2020, 12:53 AM), https://twitter.com/realDonaldTrump/status/1266231100780744704.

51. *See* Katherine Faulders, *Trump Signs Executive Order Targeting Social Media Companies*, ABC 7 NEWS (May 29, 2020), https://abc7news.com/trump-signs-executive-order-targeting-social-media-companies/6218710 [https://perma.cc/V28A-LWY7].

52. *See* Michael Wines, *'Looting' Comment from Trump Dates back to Racial Unrest of the 1960s*, N.Y. TIMES (May 29, 2020), https://www.nytimes.com/2020/05/29/us/looting-starts-shooting-starts.html; Beer, *supra* note 48.

53. *See infra* Part II (arguing that internet platforms' stated policies and practices do not adequately explain how they operationalize nonpartisanship as a principle of content moderation).

This Article examines whether internet platforms should adopt, as a matter of best practices, a principle of nonpartisanship in their moderation of content posted by political candidates, with safeguards designed to enforce that principle. Republican lawmakers have asserted that a principle of "political neutrality" exists or should exist under Section 230 of the CDA, which provides internet platforms broad immunity from civil liability.[54] But the conservative editorial board of the *Wall Street Journal* contends that "[t]rying to enforce online 'neutrality' is a fool's errand."[55] Legal scholars contend that Section 230 contains no such requirement of neutrality; the law was meant to encourage internet platforms to engage in moderation of content that is "objectionable, whether or not such material is constitutionally protected."[56]

Neither side is right. Section 230(c)(2)'s meaning of what constitutes "good faith" moderation of "objectionable" material is somewhat unclear.[57] The statute doesn't define either term. The case law, which is sparse and conflicting, has not addressed whether a content moderation decision that is based on partisanship or the political affiliation of the user can qualify as "good faith" moderation of "otherwise objectionable" material.[58] Moreover, some courts have misread Section 230(c)(1), rendering (c)(2) mere surplusage. But a careful reading of Section 230's text shows that an internet platform's removal of content is subject to Section 230(c)(2), not (c)(1). Under Section 230(c)(2), content moderation based solely on the user's political affiliation or party lacks "good faith" because it is not based on anything "objectionable" in the "material." Purely partisan content moderation does not receive any immunity under Section 230.

But, instead of amending Section 230 and imposing greater regulation of the internet, this Article proposes that internet platforms

---

54. Elizabeth Nolan Brown, *Section 230 Is the Internet's First Amendment. Now Both Republicans and Democrats Want to Take It Away.*, REASON (July 29, 2019, 8:01 AM), https://reason.com/2019/07/29/section-230-is-the-internets-first-amendment-now-both-republicans-and-democrats-want-to-take-it-away/printer.

55. Editorial, *The Twitter Fairness Doctrine*, WALL ST. J. (May 28, 2020, 7:23 PM), https://www.wsj.com/articles/the-twitter-fairness-doctrine-11590708199.

56. 47 U.S.C. § 230(c)(2).

57. *See* id.

58. *See infra* Part I (stating that the paucity of case law leaves open the question whether, under Section 230(c)(2), an internet platform does not act in good faith by moderating content it deems "objectionable" because of (1) the content's political viewpoint or (2) the user's political affiliation).

voluntarily adopt, as a matter of best practices, a limited principle of nonpartisanship as a part of their content moderation of politicians. The Article offers a model framework for nonpartisan content moderation (NCM) to implement this important principle.

Part I situates the question of nonpartisanship of internet platforms in moderating politicians' content within the burgeoning literature on online governance. Surprisingly little has been written about the issue of nonpartisan content moderation, despite the contentious public debate over the purported anti-conservative bias of social media companies.[59] Part I analyzes the text of Section 230 and shows how the Ninth Circuit and other courts have misread the provision and the relationship between its two subsections (c)(1) and (c)(2). Under the proper reading, claims based on an internet platform's *publication* of objectionable user content falls within (c)(1), whereas claims based on the *removal* of user content fall within (c)(2), which requires the internet platform to moderate in "good faith." Under the correct interpretation, content moderation based solely on the third party's political affiliation or party lacks "good faith" and falls outside of Section 230 immunity. Part I summarizes the efforts by Republican lawmakers to amend Section 230 to make politically biased moderation a disqualification from its immunity. Conducting an independent review of the community standards of the major internet platforms and their current policies with respect to nonpartisanship or bias, Part II then shows their deficiencies in explaining how nonpartisanship is ensured.

Part III proposes that internet platforms adopt, as a matter of best practices, a principle of nonpartisanship with respect to moderation of content by political candidates and officials holding public office in the United States. As a matter of best practices, internet platforms should publicly commit to a transparent framework that is designed to moderate content of political candidates and political ads in a nonpartisan matter. This approach allows the necessary flexibility to accommodate the various types of internet service providers (ISP) and the fast-changing high-tech industry, which is beset with ever-evolving, large-scale problems, including malevolent, coordinated, inauthentic behavior and foreign interference with elections. Yet, at

---

59. Notable exceptions include Anupam Chander & Vivek Krishnamurthy, *The Myth of Platform Neutrality*, 2 GEO. L. TECH. REV. 400 (2018), and Frank Pasquale, *Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power*, 17 THEORETICAL INQUIRIES L. 487 (2016).

the same time, this approach sets forth a transparent NCM framework that companies can tailor to their platforms. This Part defends the approach as better than legislation or the current bills to amend or repeal Section 230.

Part IV then outlines a model procedural framework of three levels of double-blind review for internet platforms to enforce nonpartisanship in content moderation of political candidates and officials. The model NCM framework includes several key elements: (1) a defense of selective enforcement that can be raised on appeal by a political candidate whose content has been moderated, (2) the inclusion of a public advocate as amicus curiae to represent the interests of the public during the appeal and a civil rights advocate as amicus curiae to represent positions protecting civil rights, and (3) review of challenges to both removal *and* non-removal of content of political candidates and public officials. The model framework is not intended as the exclusive way of designing a system to protect nonpartisanship in content moderation. Instead, it provides an example of a process that contains checks and balances to protect content moderation from political bias or partisanship. Other frameworks are encouraged. Part V addresses concerns with the proposal.

This Article should not be taken as an indictment charging that the internet platforms are politically biased in their content moderation. Without access to their internal decisions and undisclosed standards and procedures, there is no way to evaluate the allegation. But, therein lies a problem: internet platforms should be more transparent by explicitly recognizing in their community standards a principle of nonpartisanship in the content moderation of political candidates, and disclosing to the public the safeguards and procedures designed to ensure nonpartisanship.

I. ONLINE GOVERNANCE AND THE CONTROVERSIES OVER ELECTION INTERFERENCE, VOTER SUPPRESSION, AND PERCEIVED POLITICAL BIAS OF INTERNET PLATFORMS

Part I discusses the charge that Facebook, Twitter, and other internet platforms are politically biased in how they moderate user content.[60]

60. Content moderation is a thankless task. It requires the deployment of many staff along with the constant development of sophisticated algorithmic review to counter bots and coordinated attacks. Yet internet platforms are rarely praised for content moderation—not even for keeping horrible or illegal content (e.g., child

This allegation is best understood within the larger context of the platforms' online governance, and their recent efforts to combat election interference and coordinated campaigns to depress Black voters. The Part clears up the confusion over Section 230 and the misreading by some courts in conflating its two subsections (c)(1) and (c)(2) to both apply to an internet platform's decisions to remove user content. Under the correct interpretation, only (c)(2) applies to these decisions—and it requires that such decisions to remove "otherwise objectionable" material be made in "good faith." However, a decision to remove content based solely on the identity of the user, such as the person's political affiliation, falls outside of Section 230(c)(2) immunity. Section 230 does not protect such purely partisan removal of content.

A. Online Governance by Internet Platforms

Scholars, policymakers, and theorists have recognized how large internet platforms, such as Facebook, Google, Twitter, and YouTube, govern large swaths of the internet—billions of people—by the rules they set and the practices they impose on a wide range of issues from content moderation, free expression, surveillance, privacy, adult material—and even capitalism.[61] Just as "code is law,"[62] platforms are governments.[63] Platforms govern what people do online in ways far more extensive and direct than most national governments.[64] They

---

pornography, terrorist recruitment, revenge porn, and coordinated foreign attacks) offline.

61. *See, e.g.*, SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 93–96 (2019); Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV. 1149, 1180–81 (2018); Julie E. Cohen, *Law for the Platform Economy*, 51 U.C. DAVIS L. REV. 133, 140–53 (2017); Anupam Chander, *Facebookistan*, 90 N.C. L. REV. 1807 (2012); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1662–64 (2018); Alan Z. Rozenshtein, *Surveillance Intermediaries*, 70 STAN. L. REV. 99, 176–81 (2018).

62. *See* LAWRENCE LESSIG, CODE: VERSION 2.0 5 (2006); Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553, 555 (1998).

63. *See* Kristen E. Eichensehr, *Digital Switzerlands*, 167 U. PA. L. REV. 665, 673 (2019).

64. *See* REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM 165 (2012) ("[O]ur ability to use these platforms effectively depends on several key factors that are controlled most directly by the new digital sovereigns: they control who knows what about our identities under what circumstances; our access to information; our ability to transmit and share information

act as legislatures enacting broad regulations through their online standards, such as their community standards for content moderation.[65] They act as executives in enforcing those standards.[66] They act as agencies in establishing rules implementing broad legal concepts, such as the European Union's (EU) right to be forgotten.[67] They act as courts in adjudicating and deciding private disputes, such as over copyright claims.[68] As Rebecca MacKinnon identified in 2012, Facebook, Google, and other internet platforms operate like sovereign countries—or even kingdoms, a "Facebookistan," a "Googledom," and "Twitterverse."[69] The sovereigns analogy has even been pushed to advance the notion that internet platforms are or should be neutral akin to "digital Switzerlands."[70]

Though internet platforms govern billions of people every day, the platforms typically do not have a single founding document like the Constitution to establish the institutional structures, with checks and balances, to oversee or limit their governance of people, much less to recognize protected individual rights of their users.[71] Of course, corporate structures are set forth in articles of incorporation and bylaws, and user policies are described in the terms of service agreements and community standards. Yet, unlike the Constitution, individuals have few rights under the terms of service agreements and community standards, which the internet platforms can—and do—change at will.[72] Internet platforms typically include in their terms of

---

publicly and privately; and even whom and what we can know."); Chander, *supra* note 61, at 1815.

65. *See* Klonick, *supra* note 61, at 1631–32.

66. *See id.* at 1647.

67. *See* Edward Lee, *Recognizing Rights in Real Time: The Role of Google in the EU Right to Be Forgotten*, 49 U.C. DAVIS L. REV. 1017, 1066–72 (2016); Rory Van Loo, *The New Gatekeepers: Private Firms as Public Enforcers*, 106 VA. L. REV. 467, 496–97 (2020).

68. *See* Daphne Keller, *Facebook Restricts Speech by Popular Demand*, ATLANTIC (Sept. 22, 2019), https://www.theatlantic.com/ideas/archive/2019/09/facebook-restricts-free-speech-popular-demand/598462 ("This past week, with some fanfare, Facebook announced its own version of the Supreme Court: a 40-member board that will make final decisions about user posts that Facebook has taken down."); *see also* Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. ON REG. 547, 551–52 (2016).

69. MACKINNON, *supra* note 64, at 149.

70. *See* Eichensehr, *supra* note 63, at 696.

71. *See* Keller, *supra* note 68.

72. *See, e.g.*, *Terms of Service*, FACEBOOK, https://www.facebook.com/legal/terms/update/draft2?CMS_BRANCH_ID=1534594943262990 [https://perma.cc/6D88-EFJ5] ("Unless otherwise required by law, we will notify you before we make changes to these Terms and give you an opportunity to review them before they go into effect. Once

service agreements the power to "suspend or terminate your account or cease providing you with all or part of the Services at any time for any or no reason."[73]

Of course, internet platforms are not state actors.[74] In the United States, they are not bound by the constitutional limitations, such as the First Amendment and separation of powers, which circumscribe what the federal government can do.[75] So, one may wonder why even speak of "governance," a "constitution," "due process," or "separation of powers" when discussing internet platforms.

The reason is power—and the potential for abuse of power. Internet platforms wield enormous, often unfettered power that affects people's online existence—their free speech, their privacy, their political protest, their livelihoods, etc. For example, a business delisted and unfindable on Google does not exist, practically speaking.[76] A political candidate whose account social media companies have suspended would have virtually no chance of reaching voters online, a problem especially harmful to the political chances of a grassroots or unknown candidate.[77] The degree to which internet platforms govern people's online lives raises a set of profound questions. How should internet platforms govern? Should they incorporate practices, standards, and safeguards that exist in public law? Should national governments in turn regulate how platforms govern, or should they be left to their own devices as private corporations, especially given the long-held concern about

---

any updated Terms are in effect, you will be bound by them if you continue to use our Products.").

73. *Twitter Terms of Service*, TWITTER, https://twitter.com/en/tos#:~:text=General-,1.,old%2C%20to%20use%20the%20Services [https://perma.cc/4GW3-WY7F].

74. *See* Klonick, *supra* note 61, at 1611 (discussing *Packingham v. North Carolina*, 137 S. Ct. 1730 (2017)).

75. *See id.*

76. *See, e.g.*, Eric Goldman, *Google Must Answer Lawsuit for Manually Removing Websites from Its Search Index*, FORBES (May 17, 2016, 11:07 AM), https://www.forbes.com/sites/ericgoldman/2016/05/17/google-must-answer-lawsuit-for-manually-removing-websites-from-its-search-index/#64f6878a368b [https://perma.cc/E696-KL29].

77. *Cf.* Niam Yaraghi, *Twitter's Ban on Political Advertisements Hurts Our Democracy*, BROOKINGS INST. (Jan. 8, 2020), https://www.brookings.edu/blog/techtank/2020/01/08/twitters-ban-on-political-advertisements-hurts-our-democracy [https://perma.cc/9CPZ-SB3F] (arguing that social media can help candidates with fewer resources than "candidates with the greatest financial support from corporations and super PACs who can bankroll expensive marketing campaigns").

countries regulating—and restricting—the internet? These questions are not academic. Their answers profoundly affect free expression, elections, political protests, surveillance, and even political revolutions.[78]

This Article begins with the premise that internet platforms must think more systematically about their powers of governing people—and how they can wield those powers in ways consonant with democratic principles, including transparency, due process, and equal protection. Facebook, Twitter, and other platforms should view their responsibilities not just as profit-seeking businesses, but also as framers, like Hamilton and Madison, of the constitution of both (1) governance structures and (2) protections of individual rights on their platforms. Although these companies have already undertaken such governing responsibilities,[79] they need to do more. But it's Pollyannaish to expect corporations will transform themselves in ways that will magically solve the contentious issues of content moderation, surveillance, privacy, election interference, civil rights, etc. If national governments haven't even solved these problems offline, is it reasonable to expect Google to do so online?[80] But that is not to say people should not demand more from internet platforms. We can and should. Given how internet platforms can be weaponized to interfere with and undermine democratic elections, internet platforms must have greater accountability to the people.[81]

This Article tackles a discrete issue within that much larger endeavor: should internet platforms moderate the online content of political candidates and political campaign ads in a nonpartisan

---

78. For more on the use of social media for political protests and revolutions, see ZEYNEP TUFEKCI, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST 6 (2017), and John T. Jost et al., *How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks*, 39 ADVANCES IN POL. PSYCH. 85 (2018).

79. *See, e.g.*, *Elections Integrity: We're Focused on Serving the Public Conversation.*, TWITTER, https://about.twitter.com/en_us/advocacy/elections-integrity.html [https://perma.cc/NWB9-C2PK] (pledging to continue protecting Twitter against outside manipulation); Nick Clegg, *Welcoming the Oversight Board*, FACEBOOK (May 6, 2020), https://about.fb.com/news/2020/05/welcoming-the-oversight-board [https://perma.cc/ZN9D-QJZA].

80. *See* Keller, *supra* 68 ("But we should not fool ourselves that mimicking a few government systems familiar from grade-school civics class will make internet platforms adequate substitutes for real governments, subject to real laws and real rights-based constraints on their power.").

81. *See generally* S. Select Comm. on Intel., 116th Cong., Rep. on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2: Russia's Use of Social Media with Additional Views (Comm. Print 2019).

manner? If so, what institutional structures, checks and balances, and individual rights should be established to serve that goal? Although the issue of content moderation of political leaders and political ads is a tiny subset of content moderation on internet platforms, its relationship to political debate and elections make it an issue of double importance: as both an issue of online governance and a critical component of democratic governance related to elections.

B. Election Misinformation on Social Media in 2016 and 2020

This section explains how internet platforms were vulnerable to foreign interference and election misinformation in the 2016 U.S. election—and how they avoided a repeat of their mistakes in the 2020 U.S. election. But the increased content moderation also sparked controversy.

1. Misinformation and voter suppression during the 2016 U.S. election

The internet platforms' concern about content moderation—and how their platforms can be abused—intensified in the aftermath of the 2016 U.S. presidential election. Even before the election, BuzzFeed News conducted an October 2016 study that found "three big right-wing Facebook pages published false or misleading information 38% of the time during the period analyzed, and three large left-wing pages did so in nearly 20% of posts."[82] Shortly following the 2016 U.S. election, Zuckerberg flatly rejected the idea that misinformation on Facebook affected the outcome: "[T]he idea that fake news, of which it's a very small amount of the content, influenced the election in any way is a pretty crazy idea."[83] Later that week, under mounting criticism, Zuckerberg posted on his Facebook page the efforts Facebook was taking to stop misinformation, which he still characterized as "relatively small" on Facebook.[84]

---

82. *See* Craig Silverman et al., *Hyperpartisan Facebook Pages Are Publishing False and Misleading Information at an Alarming Rate*, BUZZFEED NEWS (Oct. 20, 2016, 12:47 PM), https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis#.pxKOlN4yp [https://perma.cc/9NHJ-JABN].

83. *See* Adrienne Jane Burke, *Facebook Influenced Election? Crazy Idea, Says Zuckerberg*, TECHONOMY (Nov. 11, 2016, 1:38 PM), https://techonomy.com/2016/11/28196 [https://perma.cc/6FBZ-24CC].

84. *See* Mark Zuckerberg, FACEBOOK (Nov. 18, 2016), https://www.facebook.com/zuck/posts/a-lot-of-you-have-asked-what-were-doing-about-misinformation-so-i-wanted-to-give/10103269806149061 [https://perma.cc/983G-RVQW].

But it soon came to light that Facebook's potential complicity was far worse than Zuckerberg admitted. In fact, Russian operatives, at the behest of the Russian government, waged a massive misinformation campaign on Facebook, Instagram, and other sites to aid Trump's candidacy and to hurt Hillary Clinton's bid.[85] The U.S. Senate Select Committee on Intelligence examined the intelligence on Russian interference in the 2016 election and issued several volumes of bipartisan reports detailing how the Russian Internet Research Agency ("IRA") "used social media to conduct an information warfare campaign designed to spread disinformation and societal division in the United States."[86] The Committee found: "[T]he IRA sought to influence the 2016 U.S. presidential election by harming Hillary Clinton's chances of success and supporting Donald Trump at the direction of the Kremlin."[87] Even further: "Russia's targeting of the 2016 U.S. presidential election was part of a broader, sophisticated, and ongoing information warfare campaign designed to sow discord in American politics and society."[88]

The Senate Intelligence Committee report is shocking. It provides numerous examples and statistics on the IRA's use of fake accounts, fake ads, and fake content on Facebook, Instagram, Twitter, Google, YouTube, Reddit, Tumblr, LinkedIn, Medium, Pinterest, and other platforms.[89] For example:

> Data provided to the Committee indicates that the IRA used 133 Instagram accounts to publish over 116,000 posts. By comparison, the IRA used Facebook pages to publish over 60,000 posts. Engagement with fellow platform users was also significantly greater on Instagram, where IRA accounts accumulated 3.3 million followers and generated 187 million total engagements. By comparison, the IRA's Facebook page audience of 3.3 million produced 76 million virtual interactions. . . . The IRA's Instagram accounts focused on both the political left and right in America, and exploited the social, political, and cultural issues most likely to incite impassioned response across the ideological spectrum. Significantly, a discernible

85. *See* S. Select Comm. on Intel., 116th Cong., Rep. on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2: Russia's Use of Social Media with Additional Views 4, 6 (Comm. Print 2019).

86. *Id.* at 3.

87. *Id.* at 4.

88. *Id.* at 5.

89. *Id.* at 43–62.

> emphasis on targeting African-Americans emerges from analysis of the IRA's Instagram activity.[90]

One of the most troubling aspects of the Russian interference in the 2016 election was its targeting of the Black community in its misinformation campaign designed to cast doubt on the relevance of the election to Black voters and Clinton's concern for Blacks, and to encourage votes for Green Party candidate Jill Stein.[91]

> The [Senate Intelligence] Committee found that no single group of Americans was targeted by IRA information operatives more than African-Americans. By far, race and related issues were the preferred target of the information warfare campaign designed to divide the country in 2016. Evidence of the IRA's overwhelming operational emphasis on race is evident in the IRA's Facebook advertisement content (over 66 percent contained a term related to race) and targeting (locational targeting was principally aimed at African-Americans in key metropolitan areas [ ]), its Facebook pages (one of the IRA's top performing pages, "Blacktivist," generated 11.2 million engagements with Facebook users), its Instagram content (five of the top 10 Instagram accounts were focused on African-American issues and audiences), its Twitter content (heavily focused on hot-button issues with racial undertones, such as the NFL kneeling protests), and its YouTube activity (96 percent of the IRA's YouTube content was targeted at racial issues and police brutality).[92]

Equally troubling is how Cambridge Analytica, a British political consulting firm working for Trump's campaign, was able to scrape personal information of eighty-seven million Facebook users without their permission.[93] Cambridge Analytica then targeted, with psychological profiling, specific impressionable voters in swing states to support Trump.[94] According to former employee and whistleblower

---

90. *Id.* at 48.

91. *See* Janell Ross, *Russia's Election Interference Exposes America's Achilles' Heel: Race*, NBC NEWS (Dec. 19, 2018, 5:16 PM), https://www.nbcnews.com/news/nbcblk/russia-s-election-interference-exposes-america-s-achilles-heel-race-n949796 [https://perma.cc/YHT5-X2UE].

92. S. Select Comm. on Intel., 116th Cong., Rep. on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2: Russia's Use of Social Media with Additional Views 6–7.

93. *See* Cecilia Kang & Sheera Frenkel, *Facebook Says Cambridge Analytica Harvested Data of up to 87 Million Users*, N.Y. TIMES (Apr. 4, 2018) https://www.nytimes.com/2018/04/04/technology/mark-zuckerberg-testify-congress.html.

94. *See* Matthew Rosenberg et al., *How Trump Consultants Exploited the Facebook Data of Millions*, N.Y. TIMES (Mar. 17, 2018), https://www.nytimes.com/2018/03/17/

Christopher Wylie, Cambridge Analytica had a program headed by Steven Bannon, who was also an adviser to Trump's campaign, to engage in voter suppression, specifically targeting Black voters.[95]

In the 2016 election, Black voter participation dropped precipitously. As Pew Research Center reported, "The [B]lack voter turnout rate declined for the first time in 20 years in a presidential election, falling to 59.6% in 2016 after reaching a record-high 66.6% in 2012. The 7-percentage-point decline from the previous presidential election is the largest on record for [B]lacks."[96]

Internet platforms were caught off-guard. They engaged in intense soul-searching after the 2016 election—and may still be continuing to do so. Updating their community standards, internet platforms implemented new policies to combat election interference, coordinated inauthentic behavior, misinformation campaigns, and voter suppression, especially of Black voters, on their platforms.[97] The controversial decisions to moderate, starting in May 2020, Trump's social media posts detailed in the Introduction can only be properly understood against this backdrop.[98] But even then, the internet platforms' moderation of Trump's posts leading up to the 2020 election was modest. Most internet platforms did not remove Trump's violating content, but instead added a label of how his

---

us/politics/cambridge-analytica-trump-campaign.html; Carole Cadwalladr, *'I Made Steve Bannon's Psychological Warfare Tool': Meet the Data War Whistleblower*, GUARDIAN (Mar. 18, 2018, 5:44 AM), https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump [https://perma.cc/8TJJ-77LN].

95. *See* Devin Coldewey, *Bannon and Cambridge Analytica Planned Suppression of Black Voters, Whistleblower Tells Senate*, TECHCRUNCH (May 16, 2018), https://techcrunch.com/2018/05/16/bannon-and-cambridge-analytica-planned-suppression-of-black-voters-whistleblower-tells-senate [https://perma.cc/M8WS-CGWG].

96. Jens Manuel Krogstad & Mark Hugo Lopez, *Black Voter Turnout Fell in 2016, even as a Record Number of Americans Cast Ballots*, PEW RSCH. CTR. (May 12, 2017), https://www.pewresearch.org/fact-tank/2017/05/12/black-voter-turnout-fell-in-2016-even-as-a-record-number-of-americans-cast-ballots [https://perma.cc/S2N6-J9K4].

97. *See, e.g.*, Guy Rosen et al., *Helping to Protect the 2020 U.S. Elections*, FACEBOOK (Oct. 21, 2019), https://about.fb.com/news/2019/10/update-on-election-integrity-efforts [https://perma.cc/UX5E-43YM]; Del Harvey & Yoel Roth, *An Update on Our Elections Integrity Work*, TWITTER (Oct. 1, 2018), https://blog.twitter.com/en_us/topics/company/2018/an-update-on-our-elections-integrity-work.html [https://perma.cc/RP7P-7GGF]; Leslie Miller, *How YouTube Supports Elections*, YOUTUBE OFF. BLOG (Feb. 3, 2020), https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html [https://perma.cc/XL45-BK7S].

98. *See supra* notes 2, 6 and accompanying text.

content violated the community standards.[99] These platforms exercised their discretion under what they call a "public interest" or "newsworthiness" exception, which allows the company to decide whether it is in the public interest to leave the violating content online for people to see, but with a label.[100]

2. Content moderation and new measures during and after the 2020 U.S. election, including suspension of President Trump's accounts

The internet platforms did a far better job in combating foreign interference in the 2020 U.S. election. The members of Election Infrastructure Government Coordinating Council Executive Committee and the members of the Election Infrastructure Sector Coordinating Council, which included federal and state officials in elections and security, as well as industry and nonprofit representatives, issued a rare joint statement after the election, describing it as "the most secure in American history."[101] They concluded: "While we know there are many unfounded claims and opportunities for misinformation about the process of our elections, we can assure you we have the utmost confidence in the security and integrity of our elections, and you should too."[102]

As the statement indicated, there still were "many unfounded claims" about the election. The source of the election misinformation was not foreign, but domestic: indeed, much from Trump and his allies.[103] On November 7, 2020, Trump declared himself the victor in several tweets.[104] According to the *Times*, from November 3 to 5, 2020,

---

99. *See supra* note 4 and accompanying text.

100. *See, e.g.*, *About Public-Interest Exceptions on Twitter*, *supra* note 49; Nick Clegg, *Facebook, Elections and Political Speech*, FACEBOOK (Sept. 24, 2019), https://about.fb.com/news/2019/09/elections-and-political-speech [https://perma.cc/MG95-NJBX].

101. *Joint Statement from Elections Infrastructure Government Coordinating Council & the Election Infrastructure Sector Coordinating Executive Committees*, CYBERSECURITY & INFRASTRUCTURE SEC. AGENCY (Nov. 12, 2020), https://www.cisa.gov/news/2020/11/12/joint-statement-elections-infrastructure-government-coordinating-council-election [https://perma.cc/29LH-6CAL].

102. *Id.*

103. *See, e.g.*, Ken Dilanian, *The Russians Have No Need to Spread Misinformation. Trump and His Allies Are Doing It for Them.*, NBC NEWS (Nov. 5, 2020, 5:15 PM), https://www.nbcnews.com/politics/2020-election/russians-have-no-need-spread-misinformation-trump-his-allies-are-n1246653 [https://perma.cc/J95A-SYCU].

104. *See President Trump Claims Victory in Series of Tweets*, WFXR (Nov. 7, 2020, 5:03 PM), https://www.wfxrtv.com/news/your-local-election-hq/president-trump-claims-victory-in-series-of-tweets [https://perma.cc/TE5Q-YVYZ].

Twitter added warning labels to thirty-eight percent of Trump's tweets; some of the labels indicated that the tweets "might be misleading about an election or other civic process."[105] The labels operated as a screen that covered the tweet and allowed the user to click "Learn more" to go to the content that disputed the claim in the tweet, or to click "View" to view the tweet.[106] Facebook added labels to Trump's posts, without the need to click through, and employed a "virality circuit-breaker" that slowed the spread of suspicious content to give more time for fact-checking.[107] Facebook removed the group "Stop the Steal," which organized based on the false claim that the election was being stolen from Trump.[108] Facebook explained: "The group was organized around the delegitimization of the election process, and we saw worrying calls for violence from some members of the group."[109]

Unfortunately, violence erupted on January 6, 2021 when pro-Trump supporters, incited by his calls to "go[] to the Capitol" and "give our Republicans, the weak ones . . . the kind of pride and boldness that they need to take back our country,"[110] attacked the Capitol during Congress's certification of Biden's election as President.[111]

---

105. Conger, *supra* note 14. Later, Twitter updated the label to indicate "Election officials have certified Joe Biden as the winner of the U.S. Presidential election." *See Twitter Updates Its Warning Labels on Political Tweets to Reflect Biden Certification*, DEADLINE (Dec. 20, 2020, 2:34 PM), https://deadline.com/2020/12/twitter-updates-warning-labels-on-tweets-1234659874 [https://perma.cc/4VYU-L3BU].

106. *See* Alex Hider, *Twitter Has Put Disclaimers on More than a Dozen Trump Tweets Since Wednesday Morning*, DENVER CHANNEL (Nov. 6, 2020, 11:45 AM), https://www.thedenverchannel.com/news/election-2020/trump-has-tweeted-32-times-since-the-polls-closed-twitter-has-applied-disclaimers-to-13-of-them.

107. *See* Tiffany C. Li, *Twitter and Facebook's Election Disinformation Efforts May Be Too Little, Too Late*, MSNBC (Nov. 11, 2020, 5:52 PM) https://www.msnbc.com/opinion/twitter-facebook-s-election-disinformation-efforts-may-be-too-little-n1247441 [https://perma.cc/RS8V-VFMN]; Kevin Roose, *On Election Day, Facebook and Twitter Did Better by Making Their Products Worse*, N.Y. TIMES (Nov. 5, 2020), https://www.nytimes.com/2020/11/05/technology/facebook-twitter-election.html.

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

109. *Id.*

110. David Z. Morris, *'We Will Never Concede': How Donald Trump Incited an Attack on America*, FORTUNE (Jan. 7. 2021, 1:45 PM), https://fortune.com/2021/01/07/trump-speech-capitol-attack-riots-pence-we-will-never-concede-maga-rally [https://perma.cc/PA56-6TDC].

111. *See How Pro-Trump Insurrectionists Broke into the U.S. Capitol*, *supra* note 16.

Trump's incitement of the insurrectionists by his baseless claims of a "stolen" election, before and after the attack on Congress, led to the internet platforms' most severe measures against Trump, including a complete, indefinite ban from posting on Facebook at least for the remaining two weeks of his presidency and Twitter's termination of Trump's personal account and freezing of the official account for the President.[112] The companies feared Trump's further use of their platforms to incite violence or insurrection, including at the inauguration of President Biden[113]—concerns that appeared justified based on "Stop the Steal" supporters' reported intention to disrupt the inauguration on January 20, 2021.[114] The fear of further insurrection and violence even led tech companies that are not involved in content moderation to act. Google removed the Parler app from its app store for Android phones, and Apple did the same for iPhones because Parler, the new social media platform popular among conservatives, had reportedly failed to moderate calls for violence or insurrection on its platform.[115] Amazon announced it would terminate its web hosting for Parler on January 10, 2021 due to violations of Amazon's rules against calls for violence.[116]

The internet platforms' and tech companies' bold actions against Trump to prevent further insurrection after the attack on the Capitol

---

112. *See* Fischer & Gold, *supra* note 19; Brian Fung, *Facebook Bans Trump from Posting for Remainder of His Term in Office*, CNN (Jan. 7, 2021, 3:37 PM), https://www.cnn.com/2021/01/07/tech/facebook-trump-restrictions/index.html [https://perma.cc/NKQ2-X7SB]; *Permanent Suspension of @realDonaldTrump*, TWITTER (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html [https://perma.cc/MUR3-DCBQ]; *Twitter Deletes New Trump Tweets on @POTUS, Suspends Campaign Account*, REUTERS (Jan. 8, 2021, 8:58 PM), https://www.reuters.com/article/us-usa-election-trump-twitter-removal/twitter-deletes-new-trump-tweets-on-potus-suspends-campaign-account-idUSKBN29E02H.

113. *See, e.g.*, Mark Zuckerberg, FACEBOOK (Jan. 7, 2021, 7:47 AM), https://www.facebook.com/zuck/posts/10112681480907401; *Permanent Suspension of @realDonaldTrump*, *supra* note 112.

114. *See* Trevor Hughes, *'It Needed to Happen': Trump Supporters Defiant After Capitol Attack, Plan to Do It Again for Biden's Inauguration*, USA TODAY (Jan. 8, 2021, 8:20 PM), https://www.usatoday.com/story/news/nation/2021/01/07/inauguration-day-violence-could-next-after-us-capitol-attack/6584582002 [https://perma.cc/KLG4-JPGD].

115. *See* Jack Nicas & Davey Alba, *Amazon, Apple and Google Cut off Parler, an App that Drew Trump Supporters*, N.Y. TIMES (Jan. 11, 2021, 11:10 AM), https://www.nytimes.com/2021/01/09/technology/apple-google-parler.html.

116. *See* John Paczkowski & Ryan Mac, *Amazon Is Booting Parler off of Its Web Hosting Service*, BUZZFEED NEWS (Jan. 9, 2021, 10:08 PM), https://www.buzzfeednews.com/article/johnpaczkowski/amazon-parler-aws [https://perma.cc/DW8E-K9UV].

were roundly praised and roundly criticized.[117] Legal scholars and attorneys from the ACLU and the Knight First Amendment Institute had various, mixed views on Twitter's eventual decision to permanently suspend Trump's account—also called "deplatforming."[118] Some were "uneasy about the developments, which underscored the enormous power of a handful of social media companies that are largely insulated from accountability and may change positions on what speech is acceptable as executives come and go."[119] This unease cut in both directions, however, with some critics wanting more content moderation and others wanting less.[120] And the controversial decision further inflamed the political debate over Section 230. As Senator Lindsey Graham tweeted: "I'm more determined than ever to strip Section 230 protections from Big Tech (Twitter) that let them be immune from lawsuits."[121]

The January 6, 2021 insurrection and its aftermath will be studied for years to come. One thing is clear: in the current political climate, it is impossible for internet platforms to moderate the content of political candidates without sparking controversy and distrust in some sectors. Indeed, the controversy following the January 6th insurrection was just another example of the ongoing dilemma that internet platforms face. When they moderate content affecting or

---

117. *See, e.g.*, Mike Isaac & Kate Conger, *Facebook Bars Trump Through End of His Term*, N.Y. TIMES (Jan. 8, 2021), https://www.nytimes.com/2021/01/07/technology/facebook-trump-ban.html.

118. *See* Adam Liptak, *Can Twitter Legally Bar Trump? The First Amendment Says Yes*, N.Y. TIMES (Jan. 9, 2021), https://www.nytimes.com/2021/01/09/us/first-amendment-free-speech.html; Kate Conger & Mike Isaac, *Twitter Permanently Bans Trump, Capping Online Revolt*, N.Y. TIMES (Jan. 12, 2021), https://www.nytimes.com/2021/01/08/technology/twitter-trump-suspended.html.

119. *Id.*; *see* Evelyn Douek, *Trump Is Banned. Who Is Next?*, ATLANTIC (Jan. 9, 2021), https://www.theatlantic.com/ideas/archive/2021/01/trump-is-banned-who-is-next/617622. *But see* Kara Swisher, *It's Time for Social-Media Platforms to Permanently Ban Trump*, INTELLIGENCER (Jan. 7, 2021), https://nymag.com/intelligencer/2021/01/its-time-for-social-media-platforms-to-ban-trump-forever.html.

120. *Compare* Ian Sherr, *Trump Showed Facebook, Twitter, YouTube Can't Moderate Their Platforms. That Needs to Change.*, CNET (Jan. 11, 2021, 2:16 PM), https://www.cnet.com/news/trump-showed-facebook-twitter-youtube-cant-moderate-their-platforms-we-need-change [https://perma.cc/XG5B-SX8B], *with* Ward Jolles, *SC's Graham Says Twitter Made 'Serious Mistake' in Banning Trump*, ABC 15 NEWS (Jan. 9, 2021), https://wpde.com/news/local/scs-graham-says-twitter-made-serious-mistake-in-banning-trump [https://perma.cc/M4QH-V6NZ].

121. Lindsey Graham (@LindseyGrahamSC), TWITTER (Jan. 8, 2021, 8:15 PM), https://twitter.com/LindseyGrahamSC/status/1347713461246169089.

related to electoral politics, the candidate or party negatively affected is likely to disagree vehemently.

For example, before the election, in October 2020, Twitter blocked a controversial *New York Post* article, which reported a Ukrainian business man's alleged emails to Hunter Biden indicating a putative meeting with Joe Biden and seeking to use Hunter Biden's "influence."[122] Twitter initially blocked links to the article as a violation of its "hacked content" policy because the emails allegedly came from a laptop repairperson's unauthorized access to the laptop.[123] But, after backlash from Trump and conservatives, Twitter quickly reversed its decision and revised its "hacked content" policy to apply only to instances in which the hacked content comes directly from the hackers or people working with them.[124] Facebook allowed links to the *New York Post* article but downgraded its prominence on users' news feeds on Facebook pending fact-checking.[125] A week before the election, Republican lawmakers on the Senate Committee on Commerce, Science, and Transportation grilled the CEOs of Twitter and Facebook about their alleged censorship of the *New York Post* article and other content.[126] Two weeks after the election, Republicans on the Senate Judiciary Committee did the same.[127] Lawmakers threatened to saddle the internet platforms with regulation.

---

122. *See* Kate Conger & Mike Isaac, *In Reversal, Twitter Is No Longer Blocking New York Post Article*, N.Y. TIMES (Dec. 28, 2020), https://www.nytimes.com/2020/10/16/technology/twitter-new-york-post.html.

123. *See* Arjun Kharpal, *Twitter Changes Hacked Material Policy After Backlash over Blocking NYPost Story About Biden's Son*, CNBC (Oct. 16, 2020, 12:50 PM), https://www.cnbc.com/2020/10/16/twitter-changes-hacked-material-policy-after-blocking-posts-biden-story.html [https://perma.cc/6L2H-8T3J]; Camille Caldera, *Fact Check: Laptop Repairman at Center of Biden Saga Is Alive*, USA TODAY (Dec. 15, 2020, 8:40 PM), https://www.usatoday.com/story/news/factcheck/2020/12/15/fact-check-laptop-repairman-center-hunter-biden-saga-alive/3905393001 [https://perma.cc/RLX9-SGR2] .

124. Kharpal, *supra* note 123.

125. *See* Shannon Bond, *Facebook and Twitter Limit Sharing 'New York Post' Story About Joe Biden*, NPR (Oct. 14, 2020, 9:14 PM), https://www.npr.org/2020/10/14/923766097/facebook-and-twitter-limit-sharing-new-york-post-story-about-joe-biden [https://perma.cc/6WJ7-PRMV].

126. *See At Hearing, Republicans Accuse Zuckerberg and Dorsey of Censorship*, N.Y. TIMES (Oct. 28, 2020, 6:28 PM), https://www.nytimes.com/live/2020/10/28/technology/tech-hearing.

127. *See Zuckerberg and Dorsey Face Harsh Questioning from Lawmakers*, N.Y. TIMES (Jan. 6, 2021, 7:11 PM), https://www.nytimes.com/live/2020/11/17/technology/twitter-facebook-hearings.

C. Section 230 of the Communications Decency Act

At the center of the debate over content moderation is Section 230, which provides immunity to internet platforms—but the scope of immunity is now contested. Republican lawmakers have proposed several bills to amend Section 230 in various ways to stop the perceived political bias by internet platforms against Trump and conservatives. Facebook also faces charges of favoritism to President Trump and conservatives on Facebook. Before examining these bills, it is important to understand the morass of case law interpreting—and in some cases misinterpreting—the provision. This section summarizes Section 230 and the circuit split, at least in dicta, in how to interpret the relationship between Section 230(c)(1) and (c)(2). Starting with the Ninth Circuit's decision in *Barnes v. Yahoo!, Inc.*,[128] some courts have misread the two provisions to allow immunity for decisions to remove content under (c)(1).[129] This Article offers and defends the correct interpretation—using a straightforward publication test for Section 230(c)(1)—which does not render the two subsections redundant as some courts do. Under this interpretation, civil claims based on an internet platform's *publication* of third-party content are potentially barred under Section 230(c)(1), while civil claims based on a platform's *removal* of such content are potentially barred under Section 230(c)(2). Thus, contrary to what some district courts have allowed, an internet platform's decision to remove or restrict access to third-party content cannot be protected under Section 230(c)(1). The only provision that applies to such removal is Section 230(c)(2). And, under that subsection, an internet platform's alleged political bias in removing content of a user can be disqualifying of civil immunity if the platform lacked "good faith" in the decision to remove "otherwise objectionable" material.

1. The confusion and controversy over Section 230

Section 230 is a source of great controversy and confusion. When the *New York Times* tried to explain it in an article on the front page of its business section, it fundamentally misstated the law in its headline, asserting that Section 230 protected hate speech.[130] The *Times* issued

128. 570 F.3d 1096 (9th Cir. 2009).

129. *Id.* at 1100–01.

130. *See* Mike Masnick, *NY Times Joins Lots of Other Media Sites in Totally and Completely Misrepresenting Section 230*, TECHDIRT (Aug. 7, 2019, 9:34 AM), https://www.techdirt.com/articles/20190806/20524742733/ny-times-joins-lots-

one of the most embarrassing corrections perhaps in the history of journalism: "An earlier version of this article incorrectly described the law that protects hate speech on the internet. The First Amendment, not Section 230 of the Communications Decency Act, protects it."[131] Even that correction left out the most important point: Section 230(c)(2) protects internet companies' ability to moderate or remove third-party content that is "objectionable," including hate speech. A day later, the *Times* published an op-ed by Jonathan Taplin that also "misstated the law containing a provision providing safe haven to social media platforms": "It is the Communications Decency Act, not the Digital Millennium Copyright Act."[132] Going for the trifecta of embarrassing errors, the *Times* issued the same correction for an article by Andrew Marantz that had the exact same mistake in confusing the Digital Millennium Copyright Act[133] (DMCA) safe harbor with the CDA immunity.[134]

The *Times* wasn't the only prominent news source to misstate Section 230. CNN published an article that erroneously described the law,[135] while both the *Washington Post* and the *Wall Street Journal* published op-eds by conservatives Charlie Kirk and Dennis Prager that legal commentators said were incorrect.[136] Sarah Jeong, a lawyer and member of the *Times* editorial board, wrote an op-ed refuting the view expressed by Republican lawmakers that Section 230 requires "a

---

other-media-sites-totally-completely-misrepresenting-section-230.shtml [https://perma.cc/ZJ68-B7VD].

131. Daisuke Wakabayashi, *Legal Shield for Websites Rattles Under Onslaught of Hate Speech*, N.Y. TIMES (Aug. 6, 2019), https://www.nytimes.com/2019/08/06/technology/section-230-hate-speech.html.

132. Jonathan Taplin, *How to Force 8Chan, Reddit and Others to Clean up*, N.Y. TIMES (Aug. 7, 2019), https://www.nytimes.com/2019/08/07/opinion/8chan-reddit-youtube-el-paso.html#click=https://t.co/pUG8F02xnj.

133. Pub. L. No. 105-304, 112 Stat. 2860 (1998).

134. *See* Andrew Marantz, Opinion, *Free Speech Is Killing Us*, N.Y. TIMES (Oct. 4, 2019), https://www.nytimes.com/2019/10/04/opinion/sunday/free-speech-social-media-violence.html.

135. *See* Brian Fung, *White House Proposal Would Have FCC and FTC Police Alleged Social Media Censorship*, CNN (Aug. 10, 2019, 8:15 AM), https://www.cnn.com/2019/08/09/tech/white-house-social-media-executive-order-fcc-ftc/index.html [https://perma.cc/SA3Y-YPUD] ("Correction: An earlier version of this story incorrectly described what content internet companies may be liable for under Section 230 of the Communications Decency Act.").

136. *See* Matthew Feeney, WSJ, WaPo, NYT *Spread False Internet Law Claims*, CATO INST. (Aug. 7, 2019, 3:24 PM), https://www.cato.org/blog/newspapers-are-spreading-section-230-misinformation [https://perma.cc/GKZ3-SGMC].

neutral public forum" and "political neutrality" to obtain Section 230's immunity.[137] Jeong contended that this political neutrality interpretation of Section 230 is a "myth . . . with no basis in law or even legislative intent."[138] Jeff Kosseff, who wrote a notable book on the history of Section 230, took the same view:

> Misunderstandings of Section 230's history already have framed the current debate, including claims that Section 230 applies only to "neutral platforms" and assumptions that Congress passed the statute to censor speech through private companies. In reality, Congress passed Section 230 so that platforms could choose not to be neutral and to moderate content based on the demands of their users (rather than regulators or judges).[139]

In testimony about the proposed Platform Accountability and Consumer Transparency (PACT) Act[140] before a Senate committee, Chris Cox, the former U.S. representative and Republican who co-sponsored the bill to enact Section 230 in 1996, agreed:

> Section 230 does not require political neutrality, and was never intended to do so. Were it otherwise, to use an obvious example, neither the Democratic National Committee nor the Republican National Committee websites would pass a political neutrality test. Government-compelled speech is not the way to ensure diverse viewpoints. Permitting websites to choose their own viewpoints is. Websites that choose to be politically neutral, and hold themselves out as such, can be held to this standard. When an internet platform promises its customers—through its advertising, published community standards, and terms of service—that its content moderation policy is politically neutral, then that promise can be enforced both by the government and civil litigants under existing federal and state laws. This is far different than a mandate of political

---

137. Sarah Jeong, *Politicians Want to Change the Internet's Most Important Law. They Should Read It First.*, N.Y. TIMES (July 26, 2019), https://www.nytimes.com/2019/07/26/opinion/section-230-political-neutrality.html#click=https://t.co/tLqhw3KfNm.

138. *Id.*

139. Jeff Kosseff, *What's in a Name? Quite a Bit, if You're Talking About Section 230*, LAWFARE (Dec. 19, 2019, 1:28 PM), https://www.lawfareblog.com/whats-name-quite-bit-if-youre-talking-about-section-230 [https://perma.cc/S649-HLSP]; *see also The PACT Act and Section 230: The Impact of the Law that Helped Create the Internet and an Examination of Proposed Reforms for Today's Online World: Hearing Before the Subcomm. on Commc'ns, Tech., Innovation, & the Internet*, 116th Cong. 11 (2020) (testimony of Jeff Kosseff, Assistant Professor, Cyber Sci. Dep't, U.S. Naval Acad.) [hereinafter *PACT Act Hearings*].

140. S. 4066, 116th Cong. (2020).

> neutrality, with the judgment of what is and is not "neutral" placed in the hands of political appointees in Washington.[141]

So, who's right? Does Section 230 require political neutrality as a prerequisite to its immunity? As explained below, I believe both sides are overstating what the text of Section 230 says or does not say—as I believe is evident by the Executive Order, DOJ report, and several bills proposed to *clarify* or *revise* Section 230, especially the meaning of "good faith."[142]

Enacted in 1996, Section 230 was drafted before social media existed. Congress had no idea of all the types or sheer scale of user-generated content that ISPs might find worrisome, whether it be disinformation attacks by foreign trolls, white supremacist propaganda, voter suppression, malicious deepfakes, or COVID-19 misinformation. Section 230's immunity is broad enough to encompass an internet company's "good faith" moderation of all these types of content if the company finds them "otherwise objectionable," even though such moderation is not viewpoint neutral. I disagree with the Republican lawmakers to the extent that they argue Section 230 requires internet platforms to be neutral public forums or maintain "viewpoint neutrality" generally for all content moderation.[143] Content moderation inevitably involves making some decisions that are not politically neutral. To take an easy case, there's no doubt Congress was contemplating that ISPs can moderate nudity and sexually explicit material under Section 230(c)(2)—such material provided a primary impetus to Senator James Exon's indecency bill to which Section 230 was proposed as an alternative approach, but was later included as an amendment to Exon's bill.[144] Yet such content moderation of nudity and sexually explicit material is not politically neutral—it discriminates against the nudist movement.[145] Moreover,

---

141. *PACT Act Hearings*, *supra* note 139, at 17 (testimony of Chris Cox, Former Member, U.S. House of Representatives).

142. *See infra* notes 343, 392, 411 and accompanying text.

143. *See* Jeong, *supra* note 137 (noting Senator Ted Cruz's statement that immunity under Section 230 is "predicate[d]" on neutrality).

144. *See* JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET 60–62 (2019). Both Exon's bill and Section 230 were enacted, but the Supreme Court struck down the indecency provisions as violating the First Amendment. *See id.* at 75–76.

145. *See* Livia Gershon, *Better Living Through Nudity*, JSTOR DAILY (Oct. 28, 2018), https://daily.jstor.org/better-living-through-nudity [https://perma.cc/4V5S-NLXT].

all content may be viewed as political in some respect.[146] Moderation of any content is, by definition, not neutral, politically or otherwise.

Yet it's a mistake to conclude that Section 230 gives internet platforms carte blanche. I disagree with Jeong, Kosseff, and Cox to the extent that they argue that the issue is foreclosed—that Section 230 precludes any argument that an internet platform's political bias in content moderation might disqualify it from "good faith" and Section 230(c)(2) immunity. To qualify for immunity from civil lawsuits, an internet platform must act in "good faith" to moderate "material" that it finds "otherwise objectionable."[147] As explained below, although courts have not squarely decided the issue, content moderation decisions lack good faith when they are based purely on the moderator's bias against or favoritism to the political party or affiliation of the user who posted the content. Such moderation wouldn't be based on what's in the "material," but simply on who posted it.

2. The circuit split over Section 230 and the misreading of Section 230(c)(1) as providing immunity for decisions to remove third-party content

Before examining Section 230(c)(2), it is important to understand its relationship with Section 230(c)(1), an issue that has bedeviled courts. One reason for all the confusion over Section 230 is that it has two different immunities in subsection (c), which is titled "Protection for 'Good Samaritan' Blocking and Screening of Offensive Material."[148] Unfortunately, in the public debate over Section 230, the provision is often discussed without differentiation or with the focus on only the first immunity, the so-called "twenty-six words that created the [i]nternet" in Kosseff's memorable phrase.[149] But focusing on Section 230(c)(1) has caused some courts and policy makers to misunderstand the section in its entirety.[150] Section 230(c) has 112

146. *See* Stanley Fish, *Is Everything Political?*, CHRON. OF HIGHER EDUC. (Mar. 29, 2002), https://www.chronicle.com/article/is-everything-political [https://perma.cc/8DAK-3394] ("Everything is political in the sense that any action we take or decision we make or conclusion we reach rests on assumptions, norms, and values not everyone would affirm.").

147. 47 U.S.C. § 230(c)(2).

148. § 230(c).

149. *See* KOSSEFF, *supra* note 144, at 2.

150. *See generally* Mont v. United States, 139 S. Ct. 1826, 1833–34 (2019) (quoting Antonin Scalia & Bryan A. Garner, READING LAW 167 (1st ed. 2012)) (noting "that 'the whole-text canon' requires consideration of 'the entire text, in view of its structure' and 'logical relation of its many parts'").

words, not just 26 words—or 129 words if the informative titles and section numbers are included. One must also consider the 25 words in Section 230(e)(3) that preempt state law: "No cause of action may be brought and no liability may be imposed under any State or local law that is inconsistent with this section."[151] This subsection makes Section 230(c)(1) an immunity from liability by preempting state law claims, similar to the express immunity in Section 230(c)(2). Thus, at a minimum, one must understand all 137 words of these subsections, plus the 25 words in the last sentence of (e)(3), to understand the two basic immunities Section 230(c) provides. Given the level of complexity among these subsections, it is perhaps not surprising there is a circuit split, at least in dicta, over the relationship between the two immunities and the proper interpretation of Section 230.

The first immunity, titled "Treatment of Publisher or Speaker," overrules the state law approach allowing online service providers to be liable for defamation based on user content posted on their bulletin boards.[152] In *Stratton Oakmont, Inc. v. Prodigy Services Co.*,[153] the state court concluded that a bulletin board operator that exercised some "editorial control," via technology and human review, over the third-party posts on the bulletin board (to make it more "family oriented"), made the operator a publisher, not a mere distributor, of the content under defamation law, thereby exposing the operator to the same standards of defamation liability that newspapers face.[154] As Justice Ain framed the issue: "In short, the critical issue to be determined by this Court is whether the foregoing evidence establishes a *prima facie* case that PRODIGY exercised sufficient editorial control over its computer bulletin boards to render it a publisher with the same responsibilities as a newspaper."[155] The court concluded Prodigy did: Prodigy Services' "content guidelines" required the removal of user content that violated its "community standards."[156] Once the internet platform engaged in any editorial control of any third-party content on its service, the platform apparently lost its status as a mere distributor of *all* content posted by third parties—regardless of whether the platform had ever reviewed

151. § 230(e)(3).
152. *See id.* § 230(c)(1).
153. No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).
154. *See id.* at *2–4.
155. *Id.* at *3.
156. *See id.* at *2.

the content in question. In effect, *Stratton Oakmont* was an all-or-nothing rule: a platform that did any moderation lost its status as a distributor for all content posted on its site. The decision created what some critics, including then-Representative Cox, viewed as a perverse result: internet platforms could be held liable if they tried to monitor and remove defamatory or objectionable material, but would not be held liable if they did nothing and had no knowledge of the offending material, a result reached in *Cubby, Inc. v. CompuServe Inc.*[157] There were other ways to analyze the role of internet platforms under the common law of defamation, but *Stratton Oakmont* arguably created a disincentive for platforms to voluntarily monitor their sites to screen objectionable content because doing so would expose them to greater liability.[158]

To avoid this result, Section 230(c)(1) creates a flat rule of immunity: "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."[159] Courts have interpreted

---

157. *See PACT Act Hearings*, *supra* note 139, at 3, 6 (testimony of Chris Cox, Former Member, U.S. House of Representatives); KOSSEFF, *supra* note 144, at 50–52. *Compare Stratton Oakmont*, 1995 WL 323710, at *5 (finding that Prodigy's choice to have editing power made it a publisher and "opened it up" to liability), *with Cubby, Inc. v. CompuServe Inc.*, 776 F. Supp. 135, 141 (S.D.N.Y. 1991) ("Because CompuServe, as a news distributor, may not be held liable if it neither knew nor had reason to know of the allegedly defamatory Rumorville statements, summary judgment in favor of CompuServe on the libel claim is granted.").

158. Eugene Volokh points out that the *Stratton Oakmont* and *Cubby* courts missed a third category of platforms under defamation law who were not treated as publishers. *See* Eugene Volokh, *47 U.S.C. § 230 and the Publisher/Distributor/Platform Distinction*, VOLOKH CONSPIRACY (May 28, 2020, 11:44 AM), https://reason.com/volokh/2020/05/28/47-u-s-c-%C2%A7-230-and-the-publisher-distributor-platform-distinction. By contrast, Benjamin Zipursky contends that the courts could have applied the traditional republication rule from defamation law in which any republication of another person's defamatory statement could expose the republisher to liability. *See* Benjamin C. Zipursky, *Online Defamation, Legal Concepts, and the Good Samaritan*, 51 VAL. U. L. REV. 1, 4–5 (2016).

159. 47 U.S.C. § 230(c)(1); *see* 141 CONG. REC. H8470 (1995) (statement of Rep. Chris Cox) ("Mr. Chairman, our amendment will do two basic things: First, it will protect computer Good Samaritans, online service providers, anyone who provides a front end to the [i]nternet, let us say, who takes steps to screen indecency and offensive material for their customers. It will protect them from taking on liability such as occurred in the Prodigy case in New York that they should not face for helping us and for helping us solve this problem. Second, it will establish as the policy of the United States that we do not wish to have content regulation by the Federal Government of what is on the [i]nternet, that we do not wish to have a

this section broadly to apply beyond defamation claims to any claim predicated on treating the online service provider as the speaker or the publisher of the allegedly offending content.[160] Even though Section 230(c)(1) is not titled as "immunity" and some courts reject that label,[161] in conjunction with Section 230(e)(3), quoted above, the section operates as an immunity by preempting civil claims that would treat online service providers as the speaker or the publisher of the user content.[162] Although this immunity is important to internet platforms as a shield from civil liability, this subsection is a red herring in the current debate over political neutrality in content moderation. Section 230(c)(1) doesn't speak to content removal, much less require "good faith" when an internet service *publishes* (rather than removes or restricts access to) content of users.

Section 230(c)(2) is the relevant subsection for the debate over political neutrality. It states:

> (2) Civil liability. No provider or user of an interactive computer service shall be held liable on account of—
>
> (A) any action voluntarily *taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable,* whether or not such material is constitutionally protected; or
>
> (B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).[163]

Importantly, like subsection (c)(1), this civil immunity applies to both providers *and users* of interactive computer services.[164] Section 230 was intended "to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate

---

Federal Computer Commission with an army of bureaucrats regulating the [i]nternet because frankly the [i]nternet has grown up to be what it is without that kind of help from the Government.").

160. *See* KOSSEFF, *supra* note 144, at 5, 92–93.

161. *See, e.g.*, City of Chicago. v. Stubhub!, Inc., 624 F.3d 363, 365–66 (7th Cir. 2010) (noting that "subsection (c)(1) does not create an 'immunity' of any kind").

162. *See* Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1100–01 (9th Cir. 2009) (explaining how Section 230(c)(1) must be read in conjunction with Section 230(e)(3), which states that "[n]o cause of action may be brought and no liability may be imposed under any State or local law that is inconsistent with this section.").

163. § 230(c)(2) (emphasis added).

164. *Id.*

online material."[165] Section 230 was also meant to encourage ISPs to be "Good Samaritans" by voluntarily engaging in content moderation and *enabling their users* (especially "parents to restrict their children's access to objectionable or inappropriate online material") the ability to do so as well.[166]

Thus, the combination of both subsections of Section 230(c) is that internet platforms can receive double immunity: first, immunity from liability, such as a defamation claim, predicated on treating them as publishers of content posted on the platforms by their users; and, second, immunity from liability for their voluntary removal or restricting access to such content—i.e., "any action voluntarily *taken in good faith* to restrict access to or availability of material *that the provider or user considers* to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable."[167]

Dictum in a Ninth Circuit decision might be interpreted as taking a contrary approach. To the extent it does, I believe it is a misreading of the statute. In *Barnes v. Yahoo!, Inc.*, a "revenge porn" case involving nude photos of the plaintiff posted by her ex-boyfriend, the Ninth Circuit held that a negligent undertaking claim against Yahoo! (for allegedly agreeing to remove the nude photos, but failing to do so) was barred by Section 230(c)(1).[168] The court viewed the tort claim for negligent undertaking as based on Yahoo!'s failure to remove published content: "*But removing content is something publishers do*, and to impose liability on the basis of such conduct necessarily involves treating the liable party as a publisher of the content it failed to remove."[169]

This syllogism attempts to provide a method for determining if a claim is barred by Section 230(c)(1) by identifying if the claim is

---

165. *Id.* § 230(b)(4).

166. *Id.*; *see PACT Act Hearings*, *supra* note 139, at 11 (testimony of Jeff Kosseff, Assistant Professor, Cyber Sci. Dep't, U.S. Naval Acad.).

167. § 230(c)(2)(A) (emphasis added); *see* KOSSEFF, *supra* note 144, at 65–66 ("Taken together, (c)(1) and (c)(2) mean that companies will not be considered to be the speakers or publishers of third-party content, and they will not lose that protection only because they delete objectionable posts or otherwise exercise good-faith efforts to moderate user content.").

168. *See* Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1105 (9th Cir. 2009); *see also* Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 367–68, 389–90 (2014) (explaining how Section 230 does not apply to criminal laws and advocating for criminal prohibitions against revenge porn).

169. *Barnes*, 570 F.3d at 1103 (emphasis added).

predicated on one of several putative functions of a publisher—what I call the functions test. This approach is apparent in the court's later discussion:

> Subsection (c)(1), by itself, shields from liability *all publication decisions*, whether to edit, *to remove*, or to post, with respect to content generated entirely by third parties. Subsection (c)(2), for its part, provides an additional shield from liability, but only for "any action voluntarily taken in good faith to restrict access to or availability of material that the provider . . . considers to be obscene . . . or otherwise objectionable."[170]

This focus on the putative functions of a publisher is similar to dictum in the Fourth Circuit's discussion of Section 230(c)(1) in the first federal appellate decision interpreting the provision, although the Fourth Circuit did not analyze the relationship with Section 230(c)(2) as the Ninth Circuit did.[171] The Ninth Circuit's language about publication involving "decisions . . . whether . . . to remove" third-party content might suggest that a platform's *decisions to remove third-party content* are protected by Section 230(c)(1)—in addition to being protected under (c)(2). But the court's suggestion was mere dictum given that *Barnes* involved a challenge to the failure to remove published content rather than its removal.

The Ninth Circuit does not appear to have decided a case applying *Barnes*'s view of Section 230(c)(1) to an internet service's removal of third-party content (as opposed to its publication) in a precedential decision. However, an unpublished Ninth Circuit decision treated MySpace's decision to remove the plaintiff's profile under Section 230(c)(1) without much analysis.[172] And, following *Barnes*, the District

---

170. *Id.* at 1105 (emphasis added) (citing § 230(c)(2)(A)).

171. *See* Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) ("Thus, lawsuits seeking to hold a service provider liable for its exercise of a publisher's traditional editorial functions—such as deciding whether to publish, withdraw, postpone or alter content—are barred."); *see also* FTC v. LeadClick Media, LLC, 838 F.3d 158, 174 (2d Cir. 2016) ("'At its core, § 230 bars 'lawsuits seeking to hold a service provider liable for its exercise of a publisher's traditional editorial functions—such as deciding whether to publish, withdraw, postpone or alter content.'" (quoting Jones v. Dirty World Ent. Recordings LLC, 755 F.3d 398, 406 (6th Cir. 2014))).

172. *See* Riggs v. MySpace, Inc., 444 F. App'x 986, 987 (9th Cir. 2011); *see also* Eric Goldman, *MySpace Quietly Won Goofy 230 Ruling in September*—Riggs v. MySpace, TECH. & MKTG. L. BLOG (Nov. 30, 2009), https://blog.ericgoldman.org/archives/2009/11/myspace_quietly.htm [https://perma.cc/EK68-H7X7] ("The court's decision is even more puzzling because 230(c)(2), which immunizes a service provider for filtering

Court for the Northern District of California has interpreted *Barnes* as treating an internet service's decisions to remove third-party content as potentially immunized under *both* (c)(1) and (c)(2).[173] For example, the district court in *Lancaster v. Alphabet*[174] ruled:

> Defendants' decision to "remov[e] content is something publishers do, and to impose liability on the basis of such conduct necessarily involves treating the liable party as a publisher." . . . Accordingly, the Court holds that § 230(c)(1) of the CDA precludes as a matter of law any claims arising from Defendants' removal of Plaintiff's videos and GRANTS the motion to dismiss to the extent that Plaintiff seeks to impose liability as a result of said removals.[175]

In a 2020 decision, the Southern District of New York agreed with this approach, while noting a split among district courts.[176]

To add to the confusion, the Seventh Circuit has suggested, in dicta, that Section 230(c)(1) should be read in a way different from *Barnes*: either as a definition (not an immunity) that clarifies that internet platforms "lose the benefit of § 230(c)(2) if it created the objectionable information" or as a bar that "forecloses any liability that depends on deeming the ISP a 'publisher' . . . while permitting the states to regulate ISPs in their capacity as intermediaries."[177] Judge Easterbrook, who advanced this alternative set of interpretations, did so to preserve a way for states to regulate ISPs and "require ISPs to protect third parties who may be injured by *material posted on their services*."[178]

---

content it subjectively deems 'objectionable,' seems to squarely cover MySpace's deletion of Riggs' account. Could the court have intended to rule for MySpace on 230(c)(2) grounds, not 230(c)(1) grounds, and just got confused?").

173. *See, e.g.*, Enhanced Athlete Inc., v. Google LLC, No. 19-cv-08260-HSG, 2020 WL 4732209, at *2–4 (N.D. Cal. Aug. 14, 2020); Ebeid v. Facebook, Inc., No. 18-cv-07030-PJH, 2019 WL 2059662, at *5 (N.D. Cal. May 9, 2019); Darnaa, LLC v. Google, Inc., No. 15-cv-03221-RMW, 2016 WL 6540452, at *7–8 (N.D. Cal. Nov. 2, 2016); Lancaster v. Alphabet Inc., No. 15-cv-05299-HSG, 2016 WL 3648608, at *3 (N.D. Cal. July 8, 2016); Sikhs for Justice "SFJ," Inc. v. Facebook, Inc., 144 F. Supp. 3d 1088, 1093–94 (N.D. Cal. 2015); Levitt v. Yelp! Inc., Nos. C-10-1321 EMC, C-10-2351 EMC, 2011 WL 5079526, at *6–7 (N.D. Cal. Oct. 26, 2011), *aff'd on other grounds*, 765 F.3d 1123 (9th Cir. 2014).

174. No. 15-cv-05299-HSG, 2016 WL 3648608 (N.D. Cal. July 8, 2016).

175. *Id.* at *3 (quoting *Barnes*, 570 F.3d at 1103).

176. *See* Domen v. Vimeo, Inc., 433 F. Supp. 3d 592, 601–04 (S.D.N.Y. 2020).

177. Doe v. GTE Corp., 347 F.3d 655, 660 (7th Cir. 2003).

178. *Id.* (emphasis added).

Let me first explain why part of Judge Easterbrook's interpretation is a misreading of Section 230 before focusing on the Ninth Circuit's misreading in *Barnes*. The idea that Section 230(c) was meant to preserve a way for state regulation of internet services runs counter to a stated goal of Section 230: "[T]o preserve the vibrant and competitive free market that presently exists for the [i]nternet and other interactive computer services, *unfettered by Federal or State regulation*."[179] Indeed, it is hard to imagine that Congress's preemption of the conflicting approaches state courts took in defamation cases before Section 230 was meant as an invitation for states to regulate providers of internet services for third-party content. In a later decision, Judge Easterbook defended his interpretation that Section 230(c)(1) should not be read as a "grant of comprehensive immunity from civil liability for content provided by a third party."[180] Judge Easterbook pointed to an internet service's liability for contributory infringement under federal copyright law, such as in *Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.*,[181] as an example to support his view.[182] But this example shows the fallacy of the interpretation. Congress added a specific exclusion of intellectual property laws from the immunities in Section 230.[183] This exclusion indicates that Congress understood that the text of Section 230 might otherwise apply to intellectual property claims. By contrast, for Section 230's effect on state laws, Congress expressly preempted "inconsistent" state law claims and liability.[184] As explained below, Judge Easterbrook's second interpretation—i.e., that Section 230(c)(1) "forecloses any liability that depends on deeming the ISP a 'publisher'"—is close to the correct interpretation.

To return to *Barnes*, the Ninth Circuit's interpretation is also a misreading of Section 230. It renders (c)(2) mere surplusage of (c)(1).[185] If nearly every decision by a provider of an interactive

179. 47 U.S.C. § 230(b)(2) (emphasis added).

180. Chi. Laws.' Comm. for C.R. Under L., Inc. v. Craigslist, Inc., 519 F.3d 666, 670 (7th Cir. 2008).

181. 545 U.S. 913 (2005).

182. *Craigslist, Inc.*, 519 F.3d at 670 (citing *Metro-Goldwyn-Mayer Studios Inc.*, 545 U.S. at 913)).

183. § 230(e)(2) ("Nothing in this section shall be construed to limit or expand any law pertaining to intellectual property.").

184. *Id.*

185. *See* TRW Inc. v. Andrews, 534 U.S. 19, 31 (2001) (detailing the canon against interpreting a statutory provision into mere surplusage).

computer service to remove third-party content falls within (c)(1), there is no need for (c)(2), which has greater requirements for such removal, including a "good faith" requirement that (c)(1) lacks. Providers of interactive computer services wouldn't have to follow (c)(2)'s "good faith" requirement to obtain immunity when removing third-party content; they would automatically receive immunity under (c)(1) simply by virtue of being providers of interactive computer services for third-party content. But this reading of (c)(1) eviscerates (c)(2). As the Middle District of Florida concluded: "But interpreting the CDA this way results in the general immunity in (c)(1) swallowing the more specific immunity in (c)(2). Subsection (c)(2) immunizes only an interactive computer service's 'actions taken in good faith.' If the publisher's motives are irrelevant and always immunized by (c)(1), then (c)(2) is unnecessary."[186]

The *Barnes* court attempts to save (c)(2) from mere surplusage by suggesting that some providers of interactive computer services might fall outside of (c)(1) immunity (e.g., if they were partly responsible for developing the content) but could qualify for (c)(2) immunity if they later restricted access to the content.[187] The court cited its divided en banc opinion in *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC*[188] as an example of a provider of interactive computer services, Roommates.com, that fell outside of (c)(1) immunity for user profiles on its site because it participated too much in the creation of the user profiles "by helping 'develop' at least 'in part' the information."[189] But, just as in *Barnes*, *Roommates.com* involved the publication of content, not its removal. So, the *Barnes* court imagines a hypothetical case of content removal to suggest how its reading of (c)(1) does not render (c)(2) mere surplusage.

This hypothetical possibility suggested by *Barnes* is too slender a reed to save (c)(2) from redundancy with (c)(1). When internet platforms, such as Twitter, Facebook, or even the early bulletin boards, remove or restrict access to third-party content, the content typically was created by their users or third parties. It would be odd, if not absurd, for Congress to enact two immunities in (c)(1) and

186. e-ventures Worldwide, LLC v. Google, Inc., No. 2:14-cv-646-FtM-PAM-CM, 2017 WL 2210029, at *3 (M.D. Fla. Feb. 8, 2017).

187. *See* Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1103 (9th Cir. 2009).

188. 521 F.3d 1157 (9th Cir. 2008) (en banc).

189. *Id.* at 1165; *see Barnes*, 570 F.3d at 1105 (citing *Roommates.com*, 521 F.3d at 1162–63).

(c)(2) that provided immunity for the *exact* same removal of third-party content. Why would Congress impose, under (c)(2), a "good faith" requirement on internet platforms only when they moderate content they created or developed, as the Ninth Circuit apparently proposes, but not when they moderate third-party content? A "good faith" requirement is only meaningful when platforms remove someone else's content. Indeed, it is hard to conceive of any lawsuit based on an internet platform's removal of its own content unless it was the platform suing itself—an absurdity on its face. If, as *Barnes* suggested, the only circumstance of content removal that (c)(2) covers that (c)(1) does not is when a provider of an interactive computer service is also the information content provider, meaning a creator or developer of the material in question, Congress could have directly stated so in (c)(2).[190] But it didn't.

3. The correct reading of Section 230(c)(1): "publisher" under the publication test

a. Treating as a "publisher" requires user content whose publication (not removal) is the basis of alleged liability

The functions test for "publisher," which treats even decisions to remove content as a function of a publisher and therefore potentially immune under Section 230(c)(1), is unmoored from the common law meaning of publisher, as well as the legislative history of Section 230 as a response to *Stratton Oakmont* and *Cubby*. Section 230(c)(1)'s inclusion of "publisher or speaker" is a nod to defamation.[191] A simpler, more direct interpretation would follow the basic common law meaning of publisher with the added identification of what Section 230(c)(1) was meant to preempt from the prior cases. Benjamin Zipursky recognized this key insight in 2016.[192] Zipursky's

190. *See generally* Azar v. Allina Health Servs., 139 S. Ct. 1804, 1813 (2019) ("So we're left with nothing but the doubtful proposition that Congress sought to accomplish in a 'surpassingly strange manner' what it could have accomplished in a much more straightforward way.").

191. *See* Blue Ridge Bank v. Veribanc, Inc., 866 F.2d 681, 686 (4th Cir. 1989) ("Public figures may not recover in a libel action absent clear and convincing proof of actual malice or of reckless disregard of the truth on *the part of the speaker or publisher of the false statements*.") (emphasis added).

192. *See* Zipursky, *supra* note 158, at 17–18.

analysis is illuminating—and well worth its own consideration—but my framing, focus, and interpretation of Section 230 are different.[193]

Although Section 230's immunity broadly bars claims beyond defamation,[194] the meaning of "publisher" is best understood under its common law meaning from defamation law.[195] To prove defamation, "the plaintiff must demonstrate that: (1) the defendant *published a defamatory statement*; (2) the defamatory statement identified the plaintiff to a third person; (3) the defamatory statement was *published to a third person*; and (4) the plaintiff's reputation suffered injury as a result of the statement."[196] In other words, a publisher is responsible for a *publication* of the statement. Without a publication, defamation law does not treat the defendant as a publisher, such as in a case in which the defendant did not make an actionable statement to a third party.[197] As the leading torts treatise instructs regarding the common-law meaning of publishers:

> Those who are in the business of making their facilities available to disseminate the writings composed, the speeches made, and the information gathered by others may also be regarded as participating to such an extent in making the books, newspapers, magazines, and information available to others as to be regarded as publishers. They are intentionally making the contents available to others, sometimes without knowing all of the contents—including the defamatory content—and sometimes without any opportunity to ascertain, in advance, the defamatory matter was to be included in the matter published. *The question is to what extent should one who is in the business of making available to the general public what another*

193. Zipursky's framing of how to understand the relationship between Section 230(c)(1) and (c)(2) focuses on tort law's recognition of an affirmative duty created based on voluntary undertakings to aid (the so-called "Good Samaritan," which Section 230(c)'s title itself references). *Id.* at 35–40. It goes beyond the scope of this Article to discuss the differences between Zipursky's article and mine.

194. *See* 47 U.S.C. § 230(e)(3) ("No cause of action may be brought and no liability may be imposed under any State or local law that is inconsistent with this section.").

195. *See* Zeran v. Am. Online, Inc., 129 F.3d 327, 331–34 (4th Cir. 1997) (applying common law principles to "publisher"). *See generally* Astoria Fed. Sav. & Loan Ass'n v. Solimino, 501 U.S. 104, 108 (1991) ("Congress is understood to legislate against a background of common-law adjudicatory principles.").

196. Cweklinksy v. Mobil Chem. Co., 837 A.2d 759, 763–64 (Conn. 2004) (emphasis added); *see* RESTATEMENT (SECOND) OF TORTS § 558 (AM. L. INST. 1977).

197. *See* Cuellar v. Walgreens Co., No. 13–00–594–CV, 2002 WL 471317, at *4 (Tex. Ct. App. Mar. 28, 2002); Brockman v. Detroit Diesel Allison Div. of Gen. Motors Corp., 366 N.E.2d 1201, 1203 (Ind. Ct. App. 1977).

> *writes or says be subject to liability for the defamatory matter that was published.* In this connection, it is necessary to classify participants into three categories: primary publishers, secondary publishers or disseminators, and those who are suppliers of equipment and facilities and are not publishers at all.[198]

Thus, under the common law, to determine if a defendant is a publisher and potentially liable based on a publication, one must examine if (1) there is a publication ("matter that was published" to a third party) and, if so, (2) whether the defendant should be considered a publisher of the publication based on the defendant's involvement in the publication. If there's no publication, the second inquiry drops out. In other words, in the absence of a publication, there's no need to consider if the defendant was a publisher under the common law.

Section 230(c)(1)'s reference to "publisher" must be interpreted against this common law background, but with the added understanding that the provision overrules the prior case law's approach to internet platforms. In short, Section 230 preempts the second inquiry above. In cases involving civil claims based on third-party content, courts do not ask "(2) whether the defendant should be considered a publisher of the publication by the defendant's involvement in the publication." Why not? Because Section 230(c)(1) precludes it: "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."[199] Section 230(c)(1) preempts examination of an internet service's involvement in the publication of third-party content online. As long as the defendant is a "provider of an interactive computer service" and the third-party content was "provided by another information content provider," the defendant will not be treated as the publisher. Instead

---

198. W. PAGE KEETON ET AL., PROSSER AND KEETON ON THE LAW OF TORTS § 113, at 803 (5th ed. 1984) (emphasis added). In the first federal appellate case interpreting Section 230, the Fourth Circuit relied on this common-law understanding of publisher. *See Zeran*, 129 F.3d at 332 (quoting W. PAGE KEETON ET AL., *supra*, § 113, at 803). In explaining how the common law treated distributors of allegedly defamatory content as publishers once they had notice of the alleged defamation, the court described how an internet service, given such notice, "must decide whether to publish, edit, or withdraw the posting," thereby assuming "the publisher role." *Id.* at 332–33. On this basis, the court interpreted Section 230(c)(1) as applying also to what the common law of defamation described as distributors. *Id.* at 330–34

199. 47 U.S.C. § 230(c)(1).

of examining whether the defendant should be considered a publisher by its involvement in the publication as the courts in *Cubby* and *Stratton Oakmont* did, courts examine under Section 230(c)(1) whether the online content was "creat[ed]" or "develop[ed]" by a third-party (i.e., "another information content provider") and not the defendant.[200] This statutory examination of who "creat[ed]" or "develop[ed]" the information in question is a different inquiry than determining who is the publisher under the common law; at least for the term "develop[ed]," courts apply a test of whether the defendant made a "material contribution" to the unlawful aspect of the content beyond its public display.[201] And courts exclude from "creat[ed]" and "develop[ed]" the traditional functions of a publisher, such as making the content publicly available; otherwise, the terms would swallow the rule against treating the internet platform as the publisher of third-party content.[202]

If courts apply this "publication" test to the meaning of "publisher" in Section 230(c)(1), courts have a straightforward inquiry—one that does not render (c)(2) mere surplusage. To determine if the claim is barred by Section 230(c)(1) under what I call the publication test, courts should examine if the claim attempts to hold the internet service liable based in part on the third-party content's *publication or public availability* on the internet service. If so (and the other conditions of (c)(1) are met, including that the content was created or developed "by *another* information content provider"[203]), the claim is barred. The publication does not have to be a formal, prima facie element of the claim, but the claim's proof of liability must involve a publication, thereby effectively treating the defendant as a publisher.[204]

---

200. § 230(f)(3) ("The term 'information content provider' means any person or entity that is responsible, in whole or in part, for the creation or development of information provided through the [i]nternet or any other interactive computer service."); *see, e.g.*, Batzel v. Smith, 333 F.3d 1018, 1031 (9th Cir. 2003) ("The 'development of information' therefore means something more substantial than merely editing portions of an e-mail and selecting material for publication."), *superseded in part by statute on other grounds as stated in*, Breazeale v. Victim Servs., Inc., 878 F.3d 759, 766–67 (9th Cir. 2017).

201. *See, e.g.*, Jones v. Dirty World Ent. Recordings LLC, 755 F.3d 398, 410–12 (6th Cir. 2014).

202. *See, e.g.*, O'Kroley v. Fastcase, Inc., 831 F.3d 352, 355 (6th Cir. 2016).

203. 47 U.S.C. §§ 203(c)(1), 203(f)(3).

204. *Cf.* Force v. Facebook, Inc., 934 F.3d 53, 64–65, 64 n.18 (2d Cir. 2019) (discussing "publisher" and relationship to publication in analyzing Section 230(c)(1).

In *Barnes*, the negligent undertaking claim did so. The claim was predicated on the revenge porn's *publication or public availability* on Yahoo!, and Yahoo!'s failure to remove the publication from the site. There's no need for the court to speculate about other putative functions of a publisher. What's essential is the existence of a *publication* of third-party content on the internet service.[205] If the case involved defamatory content instead of revenge porn, the analysis of publisher is even more obvious. Take *Stratton Oakmont*. The plaintiff claimed that the defendant internet service that operated bulletin boards should be held liable for the *publication* of a defamatory post because the defendant "was a 'publisher' of statements concerning Plaintiffs on its . . . computer bulletin board for the purposes of Plaintiffs' libel claims."[206] To determine if the defendant was a "publisher" of the publication in question, the New York state court used the common law meaning of "publisher" and focused on the level of "editorial control" the defendant internet service exercised over the bulletin board.[207] In the part that Section 230(c)(1) later overruled, the court concluded that the defendant exercised enough editorial control in the publications on its bulletin board because the defendant publicly promoted its control and it "actively utilize[ed] technology and manpower to delete notes from its computer bulletin boards on the basis of offensiveness and 'bad taste.'"[208] Section 230(c)(1) now preempts this inquiry.

Instead, under Section 230(c)(1), courts should simply ask if:

> (1) the content in question involves a publication created and developed by a third party (i.e., "another information content provider") and not by the internet service (i.e., the "provider of an interactive computer service"), and
>
> (2) does the plaintiff's claim seek to impose liability on the internet service based on the content's publication or public availability on the service ("treated as the publisher")?

If the answer to both questions is yes, the claim is barred under Section 230(c)(1). For example, the *Barnes* "revenge porn" case

---

205. *See, e.g.*, Fields v. Twitter, Inc., 200 F. Supp. 3d 964, 975 (N.D. Cal. 2016) (alteration in original) ("In defamation law, the term 'publication' means 'communication [of the defamatory matter] intentionally or by a negligent act to one other than the person defamed.'" (quoting Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1104 (9th Cir. 2009)).

206. Stratton Oakmont, Inc. v. Prodigy Servs. Co., No. 31063/94, 1995 WL 323710, at *1 (N.Y. Sup. Ct. May 24, 1995).

207. *Id.* at *3.

208. *Id.* at *4.

involved such a scenario,[209] as did *Stratton Oakmont*.[210] Alternatively, if the answer to either question is no, the claim is *not* barred under Section 230(c)(1). For example, *Roommates.com* involved a scenario in which a divided Ninth Circuit found that some of the content was not developed solely by third parties, but also involved the defendant's development, thereby making the defendant an information content provider.[211] In *Erie Insurance Co. v. Amazon.com, Inc.*,[212] the Fourth Circuit held that products liability claims against Amazon as the seller of an allegedly defective headlamp were "not based on the publication of another's speech" and therefore were not barred by Section 230(c)(1).[213]

Falling outside of (c)(1) immunity does not necessarily mean an internet platform has no immunity under Section 230. Section 230(c)(2) provides a second immunity for claims related to removing or restricting access to third-party content, as discussed below. These claims do not seek to impose liability based on a *publication* or the *public availability* of third-party content. Instead, they seek to impose liability based on the exact opposite: the absence or removal of a publication.

The publication test is the correct interpretation of Section 230. Under this approach, Section (c)(1) applies to an internet platform's decisions *not* to remove content published by third parties on the platform—meaning there is a *publication* of third-party content that is the subject of the lawsuit—as was the case in *Barnes*, while (c)(2) applies to a platform's decisions to remove or restrict third-party content.

The two subsections work in tandem, but in a complementary, not a redundant way. Claims based on the failure to remove objectionable content, such as defamation, contained in a third-party publication on an internet service fall within (c)(1). Claims based on the removal of such content fall within (c)(2). Moreover, a platform's decisions to remove some third-party content under (c)(2)—akin to what Prodigy Services did with respect to its bulletin boards—does not transform a

209. *See Barnes*, 570 F.3d at 1098–99, 1102–03.

210. *See Stratton Oakmont*, 1995 WL 323710, at *1, *4.

211. Fair Hous. Council of San Fernando Valley v. Roommates.com, 521 F.3d 1157, 1165–66 (9th Cir. 2008) (en banc).

212. 925 F.3d 135 (4th Cir. 2019).

213. *Id.* at 139–40 ("There is no claim made based on the *content of speech published* by Amazon—such as a claim that Amazon had liability as the publisher of a misrepresentation of the product or of defamatory content.").

platform into a publisher of third-party content, a general principle recognized by (c)(1). Thus, (c)(1) protects a platform's publication of third-party content, while (c)(2) gives platforms incentives to engage in some content moderation with the grant of immunity.[214]

Some passages even in *Barnes* support this reading,[215] as do some subsequent Ninth Circuit decisions.[216] Appearing to backtrack from Judge Easterbrook's first interpretation noted above, the Seventh Circuit's discussion in *Chicago Lawyers' Committee for Civil Rights Under Law, Inc. v. Craigslist, Inc.*,[217] also appears to follow the publication test.[218] To determine if Section 230(c)(1) barred a Fair Housing Act[219] claim against the website craigslist for publishing allegedly discriminatory ads posted by third parties, the Seventh Circuit simply examined whether the Fair Housing Act claim sought to impose liability on craigslist based

---

214. The *Barnes* court suggested it would be "strange" for Congress to give equal immunity for ISPs that did not remove third-party content and those that did. *Barnes*, 570 F.3d at 1105; *see also* Doe v. GTE Corp., 347 F.3d 655, 659–60 (7th Cir. 2003) (discussing, without deciding, various ways to interpret the interrelationship of Section 230(c)(1) and (c)(2)). But there's nothing strange about this approach, given Congress's stated preference in Section 230 "to preserve the vibrant and competitive free market that presently exists for the [i]nternet and other interactive computer services, unfettered by Federal or State regulation." 47 U.S.C. § 230(b)(2). In other words, Congress chose the carrot of immunity for content moderation, but did not require it.

215. *See Barnes*, 570 F.3d at 1105 ("[S]ubsection (c)(2) also protects [i]nternet service providers from liability not for publishing or speaking, but rather for actions taken to restrict access to obscene or otherwise objectionable content."); *id.* at 1105 n.11 ("It might be more straightforward to narrow the meaning of 'publisher' liability to include only affirmative acts of publication but not the refusal to remove obscene material. That path, however, is closed to us."). This passage seems to indicate that the Ninth Circuit in *Barnes* viewed decisions to remove third-party content as falling under (c)(2), not (c)(1).

216. *See, e.g.*, Kimzey v. Yelp! Inc., 836 F.3d 1263, 1268 (9th Cir. 2016) ("There is likewise no question that Kimzey's claims are premised on Yelp's publication of Sarah K's statements and star rating."); Zango, Inc. v. Kaspersky Lab, Inc., 568 F.3d 1169, 1174–75 (9th Cir. 2009) ("Section 230(c)(1) is directly aimed at the problem created by the *Stratton* decision. Section 230(c)(2)(B), on the other hand, covers actions taken to enable or make available to *others* the technical means to restrict access to objectionable material."). In an unpublished decision, the Ninth Circuit treated Facebook's decision to de-publish and then re-publish the same content the plaintiff sold to a competitor as falling within Section 230(c)(1). Fyk v. Facebook, Inc., 808 F. App'x 597 (9th Cir. 2020). The re-publication aspect of the case makes it fall within Section 230(c)(1), in my view.

217. 519 F.3d 666 (7th Cir. 2008).

218. *Id.* at 670.

219. 42 U.S.C. §§ 3601–19, 3631.

on the content's publication on craigslist.[220] The claim did, according to the court: "[O]nly in a capacity as publisher could craigslist be liable under § 3604(c) [of the Fair Housing Act]," which makes it unlawful "[t]o make, print, or publish, or cause to be made, printed, or published any notice, statement, or advertisement, with respect to the sale or rental of a dwelling that indicates any . . . discrimination based on race, color, religion, sex, handicap, familial status, or national origin."[221] Notice the Seventh Circuit didn't discuss the putative functions of publishers. Instead, the Seventh Circuit applied the publication test: the claim in question was predicated on attributing responsibility of a publication of third-party content to the internet service. In an opinion written by Judge Sutton, the Sixth Circuit took a similar approach.[222]

My proposed interpretation of Section 230 is also consistent with part of Justice Thomas's interpretation in a statement he wrote in a denial of certiorari in *Malwarebytes, Inc. v. Enigma Software Group USA, LLC*.[223] Justice Thomas criticized the functions test elaborated in *Barnes*, which has "curtailed the limits Congress placed on decisions to remove content."[224] As Justice Thomas concluded, "The decisions that broadly interpret § 230(c)(1) to protect traditional publisher functions also eviscerated the narrower liability shield Congress

---

220. *Craigslist, Inc.*, 519 F.3d at 671.

221. *Id.* at 668, 671 (quoting § 3604(c)).

222. *See, e.g.*, O'Kroley v. Fastcase, Inc., 831 F.3d 352, 355 (6th Cir. 2016) (alterations in original) ("If a website displays content that is created entirely by third parties, . . . [it] is immune from claims predicated on that content." (quoting Jones v. Dirty World Ent. Recordings LLC, 755 F.3d 398, 408 (6th Cir. 2014))).

223. *See* 141 S. Ct. 13, 14 (2020) ("Enacted at the dawn of the dot-com era, § 230 contains two subsections that protect computer service providers from some civil and criminal claims. The first is definitional. It states, '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.' § 230(c)(1). This provision ensures that a company (like an e-mail provider) can host and transmit third-party content without subjecting itself to the liability that sometimes attaches to the publisher or speaker of unlawful content. The second subsection provides direct immunity from some civil liability. It states that no computer service provider 'shall be held liable' for (A) good-faith acts to restrict access to, or remove, certain types of objectionable content; or (B) giving consumers tools to filter the same types of content. § 230(c)(2). This limited protection enables companies to create community guidelines and remove harmful content without worrying about legal reprisal.").

224. *Id.* at 17. Justice Thomas also questioned the case law holding that Section 230(c)(1)'s reference to "publishers" was meant to preclude distributor liability. *Id.* at 15–16. This issue is analyzed below.

included in the statute."[225] (As explained later, I disagree with Justice Thomas's additional suggestion that to "be treated as a publisher" does not encompass liability against distributors.)

The publication test provides a better, more straightforward interpretation of "publisher" in Section 230(c)(1) than the functions test that some courts employ. Under the publication test, decisions to remove content do not fall with (c)(1) because liability in such cases are not predicated on a publication. The removal of content involves the absence of a publication, which does not state a legal claim predicated on a publication of offending material.

b. Does "publisher" include distributors?

A final issue related to Section 230(c)(1) must be discussed because it affects the publication test discussed above. In his statement in *Malwarebytes*,[226] Justice Thomas suggested an even bigger misreading: that all courts interpreting the provision, starting in 1997 with the Fourth Circuit in *Zeran v. America Online, Inc.*,[227] have misinterpreted "publisher" to include distributors of third-party content.[228] If Justice Thomas's suggestion is correct, then Section 230(c)(1)'s immunity is far narrower than courts have uniformly recognized. Even defamation claims, such as in *Stratton Oakmont*, would be allowed against internet platforms for publishing third-party content if they had actual or constructive knowledge of the defamatory content. Internet platforms would face far greater legal liability than the current understanding of Section 230: *every* civil claim would fall outside of Section 230(c)(1) immunity as long as it has a knowledge requirement. This predicament would likely result in the unintended consequence of far more proactive, if not draconian, removal of user content by internet platforms—which critics decry as censorship. Judge J. Harvie Wilkinson III astutely foresaw this problem back in 1997, when social media didn't even exist.[229] Given the "sheer number of postings on interactive computer services," internet platforms "would have a natural incentive simply to remove messages upon notification, whether the contents were defamatory or not."[230] This problem would be exponentially

225. *Id.* at 16.
226. *Id.* at 15.
227. 129 F.3d 327 (4th Cir. 1997).
228. *See Malwarebytes*, 141 S. Ct. at 15.
229. *See Zeran*, 129 F.3d at 333.
230. *Id.* (citing Phila. Newspapers, Inc. v. Hepps, 475 U.S. 767, 777 (1986)).

worse in 2021 as the number of internet users and sheer scale of user-generated content have reached astronomical proportions.[231]

As explained below, I believe *Zeran* reached the correct interpretation of "publisher," given its common law meaning. Section 230(c)(1) bars courts from treating providers of an interactive computer service as "the publisher or speaker" of third-party content in a civil claim.[232] The uniform interpretation of "publisher" among courts that Justice Thomas questioned comes from *Zeran*, in which the Fourth Circuit rejected the plaintiff's argument that Section 230(c)(1) only preempts claims of publisher liability, but not distributor liability.[233] The plaintiff raised a negligence claim based on AOL's failure to remove alleged defamatory content even though he had provided repeated notice to AOL.[234] To avoid the Section 230(c)(1) immunity, the plaintiff argued that that it should be interpreted to allow the approach of *Cubby, Inc. v. CompuServe Inc.*, and to bar only the approach of *Stratton Oakmont*. In *Stratton Oakmont* the court held that the defendant ISP was a publisher akin to a newspaper under defamation law (therefore subject to liability without a requirement of knowledge),[235] whereas in *Cubby* a different court held that the defendant ISP was a distributor akin to a bookseller under defamation law (therefore subject to a knowledge-based requirement for liability).[236]

In rejecting the plaintiff's argument, the Fourth Circuit held that distributor liability "is merely a subset, or a species, of publisher liability, and is therefore also forcelosed by § 230."[237] In other words, "publisher" includes a distributor under the common law. The distinction between distributor and publisher liability "signifies only that different standards of liability may be applied *within* the larger

---

231. In 1997, the internet had 70 million users (or 1.7% of the world's population); in 2020, the number reached 4.8 billion users (or 62% of the world's population). *See Internet Growth Statistics*, INTERNET WORLD STATS, https://www.internetworldstats.com/emarketing.htm [https://perma.cc/MZB3-F98T]. By one estimate, Twitter alone receives 500 million tweets each day and 200 billion tweets each year. *See* David Sayce, *The Number of Tweets per Day in 2020*, DAVID SAYCE, https://www.dsayce.com/social-media/tweets-day [https://perma.cc/U4WD-6LYM].

232. 47 U.S.C. § 230(c)(1).

233. *Zeran*, 129 F.3d at 331–32.

234. *Id.* at 328.

235. Stratton Oakmont, Inc. v. Prodigy Servs. Co., No. 31063/94, 1995 WL 323710, at *5 (N.Y. Sup. Ct. May 24, 1995).

236. *See* Cubby, Inc. v. CompuServe Inc., 776 F. Supp. 135, 141 (S.D.N.Y. 1991).

237. *Zeran*, 129 F.3d at 332.

publisher category, depending on the specific type of publisher concerned."[238] "Because the publication of a statement is a necessary element in a defamation action, *only one who publishes can be subject to this form of tort liability.*"[239] In support, the Fourth Circuit quoted a passage from the influential *Prosser and Keeton on the Law of Torts*: "Those who are in the business of making their facilities available to disseminate the writings composed, the speeches made, and the information gathered by others may also be regarded as participating to such an extent in making the books, newspapers, magazines, and information available to others as *to be regarded as publishers.*"[240] And, in a passage not quoted by the Fourth Circuit, the *Prosser and Keeton* treatise offers three categories based on this recognition: "primary publishers, *secondary publishers or disseminators*, and those who are suppliers of equipment and facilities and are not publishers at all."[241]

Justice Thomas questioned *Zeran*'s interpretation, however:

> [H]ad Congress wanted to eliminate both publisher and distributor liability, it could have simply created a categorical immunity in § 230(c)(1): No provider "shall be held liable" for information provided by a third party. After all, it used that exact categorical language in the very next subsection, which governs removal of content. § 230(c)(2).[242]

Moreover, given that *Stratton Oakmont* used the terms "publisher" and "distributor"—but without Prosser and Keeton's categorization of distributors as publishers—"one might expect Congress to use the same terms *Stratton Oakmont* used" if Congress meant to preclude both approaches to liability.[243] Indeed, Congress could have just included "distributor" in the phrase "publisher or speaker." Justice Thomas's suggested alternative reading of Section 230(c)(1) has some force.[244] Some early commentary also disagreed with *Zeran*.[245]

---

238. *Id.*

239. *Id.* (emphasis added).

240. *Id.* (emphasis added) (quoting W. PAGE KEETON ET AL., *supra* note 198, § 113, at 803).

241. W. PAGE KEETON ET AL., *supra* note 198, § 113, at 803 (emphasis added).

242. Malwarebytes, Inc. v. Enigma Software Grp. USA, LLC, 141 S. Ct. 13, 16 (2020).

243. *Id.*

244. *See id.* Justice Thomas also pointed to Section 223(d) of the Communications Decency Act, 47 U.S.C. § 223(d)(1)(B)), which recognizes criminal liability for individuals who "use[] an interactive computer service" to "'knowingly . . . display' obscene material to children." *Id.* at 15 (quoting § 223(d)(1)(B)). This section of the CDA was from Senator James Exon's bill that prohibited the knowing transmission of "patently offensive" content to minors. *See* Communications Decency Act of 1996, Pub. L. 104-104, 110 Stat. 133, 133–34. The Court struck down the "patently

On the other hand, *Zeran*'s interpretation of distributor as publisher finds considerable support in the common law and the Supreme Court's own precedents. Section 230(c)(1)'s mention of "publisher or speaker" is a clear reference to the common law of defamation. The Supreme Court recognizes a canon of construction "that, absent other indication, Congress intends to incorporate the well-settled meaning of the common-law terms it uses."[246] And, as Justice Thomas recently reaffirmed, "[i]f a word is obviously transplanted from another legal source, whether the common law or other legislation, it brings the old soil with it."[247]

As *Zeran* noted, the term "publisher" under the common law of defamation applied to distributors of materials. One traditional way

---

offensive" prohibition as a violation of the First Amendment. *See* Reno v. ACLU, 521 U.S. 844, 877–79 (1997). In *Malwarebytes*, Justice Thomas concluded: "It is odd to hold, as courts have, that Congress implicitly eliminated distributor liability in the very Act in which Congress explicitly imposed it." *Malwarebytes*, 141 S. Ct. at 15.

But the answer to Justice Thomas's criticism can be found in the text of Section 230(e)(1), which states: "Nothing in this section shall be construed to impair *the enforcement of section 223* or 231 of this title, chapter 71 (relating to obscenity) or 110 (relating to sexual exploitation of children) of title 18, or any other Federal criminal statute." § 230(e)(1) (emphasis added). Section 230(e)(1) does not bar "the enforcement of section 223," including by civil enforcement against "common carriers" under Section 207, which predicates liability on "provisions of this chapter." § 207. Thus, Section 230 does not eliminate liability created by Section 223(d); instead, Section 230 expressly preserves it. Also, it is unclear whether Section 223(d) even preserves distributor liability. Section 223(d)(1) applies to an individual who knowingly "uses an interactive computer service to display" prohibited content to minors. § 223(d)(1)(B). Section 230, however, distinguishes between a "provider" and a "user" of an interactive computer service. § 230(c)(1). Arguably, one who "uses an interactive computer service" is the user, not the provider. Moreover, Section 223(d)(2) makes it a crime for anyone who "knowingly permits any telecommunications facility under such person's control to be used for an activity prohibited by paragraph (1)," but requires "the intent that it be used for such activity." § 223(d)(2). Intent is a higher standard than the knowledge requirement for distributor liability under defamation law.

245. *See, e.g.*, Susan Freiwald, *Comparative Institutional Analysis in Cyberspace: The Case of Intermediary Liability for Defamation*, 14 HARV. J.L. & TECH. 569, 637–41 (2001); David R. Sheridan, Zeran v. AOL *and the Effect of Section 230 of the Communications Decency Act upon Liability for Defamation on the Internet*, 61 ALB. L. REV. 147, 168 (1997); *see also* KOSSEFF, *supra* note 144, at 95 (arguing that the more limited reading of "publisher" was one that the Fourth Circuit "could have reasonably adopted").

246. Universal Health Servs., Inc. v. United States *ex rel.* Escobar, 136 S. Ct. 1989, 1999 (2016) (quoting Sekhar v. United States, 570 U.S. 729, 732 (2013)).

247. Stokeling v. United States, 139 S. Ct. 544, 551 (2019) (quoting Hall v. Hall, 138 S. Ct. 1118, 1128 (2018)).

of viewing the issue under the common law was that a distributor was treated as a publisher of the defamatory material if the distributor had knowledge of its defamatory content. As the Supreme Court of Minnesota explained in a 1978 decision, "[t]hose who merely deliver or transmit defamatory material previously published by another *will be considered to have published* the material only if they knew, or had reason to know, that the material was false and defamatory."[248] This understanding of distributor-as-publisher was adopted in a noteworthy 1985 federal district court decision in *Dworkin v. Hustler Magazine, Inc.*,[249] and was commonly recognized in law review articles between 1984 and 1996, when the CDA was enacted, including in discussion of how to analyze liability for bulletin boards and online dissemination.[250] Indeed, contemporaneous legal commentary even described *Cubby*, which supposedly involved distributor liability, as involving the approach to "secondary publishers."[251]

The traditional approach of distributor-as-publisher dates back to the English common law, which placed the burden on the defendant: the defendant distributor was presumptively treated as a publisher of defamatory content based on a prima facie case of defamation. As Lord Esher explained in *Emmens v. Pottle*,[252] to rebut the presumption of being a publisher, the distributor had the affirmative burden to show it lacked actual or constructive knowledge of the defamatory content to establish that the distributor "did not publish the libel."[253]

---

248. *See* Church of Scientology of Minn. v. Minn. State Med. Ass'n Found., 264 N.W.2d 152, 156 (Minn. 1978) (emphasis added).

249. *See* 611 F. Supp. 781, 785–86 (D. Wyo. 1985).

250. *See, e.g.*, Jeffrey M. Taylor, *Liability of Usenet Moderators for Defamation Published by Others: Flinging the Law of Defamation into Cyberspace*, 47 FLA. L. REV. 247, 269–70 (1995) (treating distributors as "secondary publishers" under the common law); Loftus E. Becker, Jr., *The Liability of Computer Bulletin Board Operators for Defamation Posted by Others*, 22 CONN. L. REV. 203, 215–16, 226–27 (1989) (same); Robert Charles, Note, *Computer Bulletin Boards and Defamation: Who Should Be Liable? Under What Standard?*, 2 J.L. & TECH. 121, 131 (1987) (same).

251. *See* R. Timothy Muth, *Old Doctrines on a New Frontier: Defamation and Jurisdiction in Cyberspace*, 68 WISC. LAW. 10, 12 (1995) (explaining *Cubby* as based on principle that "'[s]econdary publishers' such as libraries and bookstores are not held liable for statements in the books they offer unless they know, or have reason to know, of the existence of defamatory material in a book").

252. [1885] 16 QBD 354 (Eng.).

253. *See id.* at 357; *see also* Brief of Amici Curiae Amazon.com, Inc. et al. at 23–30, Barrett v. Rosenthal, 40 Cal. 4th 33 (No. S122953), 2004 WL 3256404 (discussing case law regarding distributor liability).

The *Emmens* doctrine (known as the defense of innocent dissemination in the United Kingdom) was influential.[254] For example, in *Vizetelly v. Mudie's Select Library, Ltd.*,[255] the Queen's Bench Division upheld the jury's finding of libel against a library (i.e., a distributor) that circulated a libelous book to the public.[256] The library claimed it had no knowledge of the contents of the book; even though the jury agreed, it rejected the library's defense.[257] On appeal, the court found sufficient evidence for the jury to reject the library's defense on the ground that the library's own negligence was the reason for their lack of knowledge of the libelous book.[258] As Lord Justice Archibald Levin Smith concluded: "That being so, they failed to do what the defendants in *Emmens v. Pottle* succeeded in doing, namely, prove that *they did not publish the libel*."[259] Thus, *Emmens* framed liability for distributors in terms of whether they should be treated as a publisher: "The question is whether, as such disseminators, *they published the libel*."[260] For a distributor to be treated as a publisher meant the distributor was liable. Conversely, if the distributor was not treated as a publisher, the distributor was not liable.

Thus, under this traditional common law understanding, to "be treated as the publisher" in Section 230(c)(1) should be interpreted to include distributors. As explained above, before the passage of Section 230 in 1996, a traditional understanding of "publisher" dating back to the English common law was that a distributor who had knowledge of defamatory content was treated as a publisher of the content (indeed, the English common law was even stronger in *presuming* a distributor was a publisher based on a prima facie case of defamation even without proof of knowledge).[261] The 1984 *Prosser and Keeton* treatise recognized this traditional understanding of distributors when classifying distributors as "secondary publishers."[262] In 1971, Prosser summarized the principle in this way: "Likewise *every one* who takes part in the publication, as in the case of the owner, editor, printer,

---

254. *See* Douglas W. Vick & Linda Macpherson, *An Opportunity Lost: The United Kingdom's Failed Reform of Defamation Law*, 49 FED. COMM. L.J. 621, 630 & n.45 (1997).
255. [1900] 2 QB 170 (Eng.).
256. *Id.* at 175.
257. *Id.* at 176.
258. *Id.*
259. *Id.* at 177 (emphasis added).
260. *Id.* at 175 (emphasis added).
261. *See supra* notes 253–60 and accompanying text.
262. *See* W. PAGE KEETON ET AL., *supra* note 198, § 113, at 803.

vendor, or even carrier of a newspaper *is charged with publication*, although so far as strict liability is concerned the responsibility of some of these has been somewhat relaxed."[263] Legal commentary in 1939 recognized the distributor-as-publisher doctrine and its English common law antecedents even more saliently:

> But it is settled by the English decisions and the few American cases on the point that *such secondary publishers who sell, rent, give, or otherwise circulate defamatory matter* originally published by a third person will be excused from liability if they show that there was no reason to know of its defamatory character.[264]

And, if there's any doubt, one need only examine the two cases that prompted Congress to enact Section 230. Even though *Cubby* and *Stratton Oakmont* did not directly discuss the distributor-as-publisher doctrine, the cases they relied on did. In its analysis of liability for distributors, *Cubby* cited[265] as its main authority *Lerman v. Chuckleberry Publishing, Inc.*,[266] which, in turn, relied on *Balabanoff v. Fossani*,[267] a New York state decision that includes as authority two decisions of other state supreme courts that recognize the traditional common-law approach in treating distributors as publishers of defamatory content unless they can show they had no knowledge of it.[268] Indeed, both cases quoted the rule from *Emmens*, the seminal English case that framed the distributor-as-publisher approach.[269] Likewise, *Stratton Oakmont* cited[270] *Cubby* and *Auvil v. CBS 60 Minutes*,[271] which framed the question of conduit liability—that the court likened to liability for book sellers (i.e., distributors)—in the following terms: "The

---

263. WILLIAM L. PROSSER, HANDBOOK OF THE LAW OF TORTS § 113, at 768–69 (4th ed. 1971) (emphasis added) (footnotes omitted).

264. Ralph E. Helper, *Libel and Slander — Privilege of "Fair and Accurate Report" of Judicial Proceedings — Non-Liability of Vendor of Newspaper*, 37 MICH. L. REV. 1335, 1336 (1939).

265. *See* Cubby, Inc. v. CompuServe Inc., 776 F. Supp. 135, 139 (S.D.N.Y. 1991).

266. 521 F. Supp. 228, 235 (S.D.N.Y. 1981) (citing Balabanoff v. Fossani, 81 N.Y.S.2d 732 (N.Y. Sup. Ct. 1948)) ("With respect to distributors, the New York courts have long held that vendors and distributors of defamatory publications are not liable if they neither know nor have reason to know of the defamation.").

267. 81 N.Y.S.2d 732 (N.Y. Sup. Ct. 1948).

268. *Id.* at 733 (citing Bowerman v. Detroit Free Press, 283 N.W. 642, 645 (Mich. 1939); Street v. Johnson, 50 N.W. 395, 396 (Wis. 1891)).

269. *See Bowerman*, 283 N.W. at 645 (citing Emmens v. Pottle, [1885] 16 QBD 354 (Eng.)); *Street*, 50 N.W. at 396 (same).

270. *See* Stratton Oakmont, Inc. v. Prodigy Servs. Co., No. 31063/94, 1995 WL 323710, at *3 (N.Y. Sup. Ct. May 24, 1995).

271. 800 F. Supp. 928 (E.D. Wash. 1992).

threshold inquiry is whether a local broadcaster who serves as a mere conduit 'republishes' by relaying an unedited feed."[272] Like "publishes," "republishes" is a term of art that here refers to the republication rule that holds liable everyone who "republishes" defamatory content.[273] For its discussion of distributor liability, *Auvil* relied on the federal district court's analysis in *Dworkin v. Hustler Magazine, Inc.*[274] Although *Auvil* cited the court's dismissal of the defamation claim against distributor Inland Empire Periodicals,[275] the *Dworkin* court had issued an earlier decision dismissing the claim against distributor Park Place Market.[276] That earlier decision in *Dworkin* relied on Prosser and Keeton's classification of distributors as "secondary publishers": "The general rule for *secondary publishers*, as stated in § 581 of the Second Restatement of Torts, is that 'one who only delivers or transmits defamatory matter published by a third person is subject to liability if, but only if, he knows or has reason to know of its defamatory character.'"[277] Thus, both *Cubby* and *Stratton Oakmont* relied on cases that recognized the distributor-as-publisher doctrine as a part of the common law. Both cases were decided under New York common law, which has used the terms "published" and "publisher" to describe when distributors and conduits are liable as publishers.[278] It's noteworthy the Court of Appeals of New York later agreed with and adopted *Zeran*'s interpretation of Section 230 and "publisher."[279] Had the *Zeran* court's understanding of distributor liability as a subset of

---

272. *Id.* at 931.

273. *See* Cianci v. New Times Publ'g Co., 639 F.2d 54, 60–61 (2d Cir. 1980).

274. 634 F. Supp. 727 (D. Wyo. 1986); *see Auvil*, 800 F. Supp. at 931–32 (quoting *Dworkin*, 634 F. Supp. at 729).

275. *See Dworkin*, 634 F. Supp. at 729.

276. *See* Dworkin v. Hustler Magazine, Inc., 611 F. Supp. 781, 785–86 (D. Wyo. 1985).

277. *Id.* at 785 (emphasis added); *see id.* at 785–86 (quoting W. PAGE KEETON ET AL., *supra* note 198, § 113, at 810).

278. *See, e.g.*, Anderson v. N.Y. Tel. Co., 320 N.E.2d 647, 647 (N.Y. 1974) (adopting the appellate court's dissenting opinion); Anderson v. N.Y. Tel. Co., 345 N.Y.S.2d 740, 751 (N.Y. App. Div. 1973) (Witmer, J., dissenting) (emphasis added) ("The telephone company cannot be liable to plaintiff under the law of defamation *unless it is held that by providing service to Jackson it 'published' his messages* and did so under circumstances such that the 'publication' was not privileged."); Wolfson v. Syracuse Newspapers, 18 N.E.2d 676, 679 (N.Y. 1939) (Rippey, J., dissenting) (per curiam) (emphasis added) (citing Vizetelly v. Mudie's Select Library, Ltd., [1900] 2 QB 170 (Eng.)) ("[I]n the case of a circulating library which bought, sold or rented books as a business for private profit, the proprietors *would be held responsible as publishers.*").

279. *See* Shiamili v. Real Estate Grp. of N.Y., Inc., 952 N.E.2d 1011, 1016–17 (N.Y. 2011).

publisher liability been contrary to New York common law, the Court of Appeals of New York presumably would have said so.

Zeran's interpretation better promotes an express policy of Section 230 "to preserve the vibrant and competitive free market that presently exists for the [i]nternet and other interactive computer services, unfettered by Federal or State regulation."[280] By contrast, interpreting Section 230 to permit civil claims based on knowledge would invite extensive—potentially conflicting—state regulation of the internet. The ramifications are sweeping: *every* civil claim would fall outside of Section 230(c)(1) immunity as long as it has a knowledge requirement. That result does not square with a "free market" that is "unfettered by Federal or State regulation." Although Justice Thomas appeared to question *Zeran*'s reliance on the express purposes of the statute,[281] a textualist reading of a statute permits consideration of a stated policy in the text of a statute.[282] *Zeran* adopts a categorical approach, which Justice Scalia recommended in interpreting the text of a statute.[283] Moreover, a categorical approach to immunity under Section 230 helps to avoid the potential dormant Commerce Clause problem in which different, possibly conflicting state law approaches create an excessive burden on interstate commerce.[284]

4. Decisions to remove or restrict access to content should be analyzed under Section 230(c)(2)'s immunity

Once we have the correct interpretation of Section 230(c)(1), the key question for content removal decisions is: does Section 230(c)(2)'s requirement of "good faith" moderation of "otherwise objectionable" material mean that internet platforms must avoid political bias or remain politically neutral in content moderation to obtain (c)(2) immunity? As

280. 47 U.S.C. § 230(b)(2).

281. Malwarebytes, Inc. v. Enigma Software Grp. USA, LLC, 141 S. Ct. 13, 15 (2020) (alteration in original) ("In reaching this conclusion, the court stressed that permitting distributor liability 'would defeat the two primary purposes of the statute,' namely, 'immuniz[ing] service providers' and encouraging 'selfregulation.'").

282. *See, e.g.*, Rapanos v. United States, 547 U.S. 715, 737 (2006) (Scalia, J.) (examining the Clean Water Act's stated policy in interpreting "waters" under the Act).

283. *See* Antonin Scalia, *The Rule of Law as a Law of Rules*, 56 U. CHI. L. REV. 1175, 1183 (1989) ("[U]nless such a statutory intent is express or clearly implied, courts properly assume that 'categorical decisions may be appropriate and individual circumstances disregarded when a case fits into a genus in which the balance characteristically tips in one direction.'" (quoting U.S. Dep't of Justice v. Reporters Comm. for Freedom of the Press, 489 U.S. 749, 775 (1989))).

284. *See generally* Pike v. Bruce Church, 397 U.S. 137, 142 (1970).

explained below, "good faith" is not defined by statute, but at least one type of situation lacks good faith: when an internet platform removes the content of a user solely based on the user's political party or affiliation. In such case, the internet platform's decision lacked good faith because it was not even based on the actual "material" or anything "otherwise objectionable" in the content, but simply based on the user's political party. Thus, if an internet platform removed the content of a politician simply because the platform wanted the politician to lose the election, that removal decision is not entitled to Section 230(c)(2) immunity.

a. "Good faith" is a subjective standard that affords leeway

Courts have interpreted "good faith" to denote a subjective standard.[285] Under a subjective standard, an internet platform has wide latitude to moderate what "*the provider . . . considers* . . . otherwise objectionable."[286] This favors interpreting Section 230 to contain no general requirement of political neutrality.

For example, Facebook considers a political or religious position taken against same-sex marriage that included an image of a same-sex couple as a violation of its community standard.[287] Facebook says it will remove the violating content.[288] Such content moderation would constitute viewpoint discrimination—against the view that same-sex marriage is wrong. Under a subjective standard of good faith, Facebook's decision falls within Section 230(c)(2). This conclusion is analogous to the one reached by the district court for the Southern

285. *See, e.g.*, Domen v. Vimeo, Inc., 433 F. Supp. 3d 592, 603–04 (S.D.N.Y. 2020) (citation omitted) ("Section 230(c)(2) is focused upon the provider's subjective intent of what is 'obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable.' That section 'does not require that the material actually be objectionable; rather, it affords protection for blocking material "that the provider or user considers to be" objectionable.'" (quoting Zango, Inc. v. Kaspersky Lab, Inc., No. C07-0807-JCC, 2007 WL 5189857, at *4 (W.D. Wash. Aug. 28, 2007))); Levitt v. Yelp! Inc., No. C-10-1321-EMC, C-10-2351-EMC, 2011 WL 5079526, at *7 (N.D. Cal. Oct. 26, 2011) ("That § 230(c)(2) expressly provides for a good faith element omitted from § 230(c)(1) indicates that Congress intended not to import a subjective intent/good faith limitation into § 230(c)(1)."), *aff'd*, 765 F.3d 1123 (9th Cir. 2014).

286. 47 U.S.C. § 230(c)(2)(A).

287. *See* Monika Bickert, *Hard Questions: Why Do You Leave up Some Posts but Take down Others?*, FACEBOOK (Apr. 24, 2018), https://about.fb.com/news/2018/04/community-standards-examples [https://perma.cc/RR5R-FYUV].

288. *Id.*

District of New York in *Domen v. Vimeo, Inc.*[289] The court held that Vimeo's removal of Church United's videos that Vimeo found violated its stated policy against content that "promote[s] Sexual Orientation Change Efforts (SOCE)" fell within Section 230(c)(2).[290] Although the plaintiffs argued that Vimeo did not act in good faith, the complaint contained "no facts to support this allegation."[291] The allegations indicated that Vimeo followed its own stated community guidelines in removing the videos.[292] Neither Facebook nor Vimeo was viewpoint neutral in these examples, but both acted in good faith according to their community guidelines to moderate "otherwise objectionable" content.

Further support for this interpretation is provided by Section 230(c)(2)'s extension of immunity to content moderation by a "user" regarding what the "*user considers* to be . . . otherwise objectionable."[293] Congress had envisioned both "the development of technologies which maximize *user control* over what information is received by individuals, families, and schools who use the [i]nternet" and "the development and utilization of blocking and filtering technologies that empower *parents* to restrict their children's access to objectionable or inappropriate online material."[294] Given that Congress wanted to promote technologies that "*maximize user control* over what information" they or their families receive, it is hard to imagine that *maximizing* user-based content moderation would entail *limiting* a user from viewpoint discrimination even though the "*user considers [it] to be . . . otherwise objectionable.*"[295]

Imagine that an internet platform gave users the ability to choose to screen out content that criticized same-sex marriage as unnatural or a sin. If a person or parents subjectively believed such content was objectionable and not appropriate for them or their children to view, presumably their decision to moderate the content they "*consider . . . objectionable*" is taken in good faith. If Congress intended the permissible scope of "good faith" content moderation of "otherwise objectionable"

289. 433 F. Supp. 3d 592, 603 (S.D.N.Y. 2020).
290. *Id.* at 603–04.
291. *Id.* at 604.
292. *Id.* ("[W]hat occurred here is that Vimeo applied its Guidelines to remove Plaintiffs' videos, since such videos violated the Guidelines.").
293. 47 U.S.C. § 230(c)(2).
294. *Id.* § 230(b)(3), (4) (emphasis added).
295. *Id.* § 230(b)(3), (c)(2)(A) (emphasis added).

material to differ between the "provider" and the "user of an interactive computer service," Congress included no words in Section 230(c)(2) recognizing such a distinction. Perhaps one can argue that the purpose of maximizing user control implies that users should have greater leeway than internet platforms in deciding what constitutes "good faith" content moderation than what constitutes "good faith" moderation by ISPs. However, this interpretation runs up against the canon of construction that recognizes a rebuttable presumption that the same term in a statute has the same meaning;[296] presumably, the canon is even stronger where, as here, it is the exact same term in the same subsection.

b. Does "otherwise objectionable" limit the permissible bases for content moderation under Section 230(c)(2)?

On the other hand, one might argue that the catchall term "otherwise objectionable" under Section 230(c)(2) should be read narrowly to limit the discretion of internet companies. Judge Conti of the Northern District of California held that the phrase "otherwise objectionable" does not apply to whatever an internet platform subjectively considers to be objectionable; instead, "otherwise objectionable" content must be "offensive."[297] Judge Conti pointed to the section heading, "Protection for 'Good Samaritan' Blocking and Screening of *Offensive* Material," as well as the *eiusdem generis* canon of construction that favors interpreting the catchall "otherwise objectionable" to be similar in kind to the immediately preceding words "obscene, lewd, lascivious, filthy, excessively violent, harassing" in the same phrase.[298] Judge Conti ruled that YouTube's proffered reason—a user improperly inflating the view count for a video on YouTube—did not constitute "otherwise objectionable" content.[299] Judge Conti cited two other district court opinions that ruled that a content moderation policy regarding pricing and cancellation

296. *See* Env't Def. v. Duke Energy Corp., 549 U.S. 561, 574 (2007).

297. *See* Song fi Inc. v. Google, Inc., 108 F. Supp. 3d 876, 883–84 (N.D. Cal. 2015) ("[T]he fact that the statute requires the user or service provider to subjectively believe the blocked or screened material is objectionable does not mean anything or everything YouTube finds subjectively objectionable is within the scope of Section 230(c).").

298. *Id.* at 882–83.

299. *See id.* at 883–84.

information in Google ads online by mobile providers[300] and eBay's content moderation of "an auction of potentially-counterfeit coins"[301] fell outside Section 230(c)(2)'s meaning of "objectionable" content.[302] This narrow interpretation of Section 230 supports the view that an internet platform cannot invoke Section 230(c)(2)'s immunity for content moderation based on an internet platform's disagreement with the political viewpoint of the user; the content itself must be "offensive" in a way similar to "obscene, lewd, lascivious, filthy, excessively violent, [or] harassing" content.[303]

However, this very narrow interpretation of Section 230(c)(2)'s meaning of "objectionable" content is open to criticism. The interpretation ignores the plain meaning of "otherwise" and runs the risk of rendering the catchall phrase "otherwise objectionable" into mere surplusage, denoting the same thing as the other listed categories.[304] The dictionary definition of "otherwise" is "in a *different* way or manner," "in *different* circumstances," and "in *other* respects."[305] Thus, a more plausible reading of the catchall term "otherwise objectionable" is that it refers to things the provider considers objectionable in a *different* way or manner, in *different* circumstances, or in *other* respects than "obscene, lewd, lascivious, filthy, excessively violent, [or] harassing" content. Given the plain meaning of "otherwise," the "*noscitur a sociis*" canon of construction of interpreting a word by the company it keeps does not apply.[306] Indeed, to read "otherwise objectionable" in narrow

300. *See* Goddard v. Google, Inc., No. C 08-2738, 2008 WL 5245490, at *6 (N.D. Cal. Dec. 17, 2008).

301. Nat'l Numismatic Certification, LLC. v. eBay, Inc., No. 6:08-CV-42-ORL-19GJK, 2008 WL 2704404, at *25 (M.D. Fla. July 8, 2008). However, the court's decision appears to render the catchall "objectionable" equivalent to and redundant of the other listed categories. *Id.* ("Accordingly, the Court concludes that 'objectionable' content must, at a minimum, involve or be similar to pornography, graphic violence, obscenity, or harassment.").

302. *See Song fi*, 108 F. Supp. 3d at 883.

303. 47 U.S.C. § 230(c)(2)(A).

304. *See, e.g.*, National Numismatic Certification, LLC. v. eBay, Inc, No. 6:08-CV-42-ORL-19GJK, 2008 WL 2704404, at *25 (M.D. Fla. July 8, 2008) ("Accordingly, the Court concludes that 'objectionable' content must, at a minimum, involve or be similar to pornography, graphic violence, obscenity, or harassment.").

305. *Otherwise*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/otherwise [https://perma.cc/V2GB-HYVX].

306. *See* Russell Motor Car Co. v. United States, 261 U.S. 514, 519–20 (1923) ("Rules of statutory construction are to be invoked as aids to the ascertainment of the meaning or application of words otherwise obscure or doubtful. They have no place, as this Court has many times held, except in the domain of ambiguity.").

fashion would contradict one of Section 230's stated goals to "maximize user control" over the content they view.[307] Users might find some content (e.g., white supremacist propaganda, or opposition to same-sex or interracial marriage) to be objectionable—or offensive, for that matter—for reasons completely unrelated to sex, violence, or harassment. As the Ninth Circuit recognized: "If the enumerated categories are not similar, they provide little or no assistance in interpreting the more general category. . . . We think that the catchall was more likely intended to encapsulate forms of unwanted online content that Congress could not identify in the 1990s."[308]

c. Removing user content based solely on identity of a user, such as the user's political party or affiliation, falls outside of Section 230(c)(2) immunity

I believe "otherwise objectionable" should be understood as a broad catchall term that does not generally require viewpoint neutrality. However, this approach doesn't give complete discretion to internet platforms. Decisions based solely on the identity of the user—such as the person's political affiliation—fall outside of Section 230(c)(2)'s immunity because they are not based on anything "otherwise objectionable" in the "material." As such, there was no action "taken in good faith to restrict access to or availability of *material* that the provider . . . considers to be . . . otherwise objectionable."[309] Instead, the action was taken against a user whose *identity*—such as political affiliation or party—the internet platform considers to be objectionable. By the plain text of Section 230(c)(2), such a removal decision does not fall within the immunity because the decision was not based on anything "otherwise objectionable" in the "material" itself.

Although no court has considered the precise issue of alleged bias against a user's political affiliation in content moderation,[310] consider

307. *See* 47 U.S.C. § 230(b)(3).

308. Enigma Software Grp. USA, LLC v. Malwarebytes, Inc., 946 F.3d 1040, 1051–52 (9th Cir. 2019).

309. § 230(c)(2) (emphasis added).

310. Other than the text of Section 230, including its several stated purposes, courts have little in the way of legislative history on what constitutes an "action . . . taken in good faith to restrict access to . . . otherwise objectionable" content. *Id.*; *see* 141 CONG. REC. H8470 (1995) (statement of Rep. Chris Cox) ("We can keep away from our children things not only prohibited by law, but prohibited by parents. That is where we should be headed, and that is what the gentleman from Oregon [Mr. Wyden] and I are doing."). Former Representative Cox's later statement before a

the controversy over Facebook's alleged blocking of access to the Facebook page of the nonprofit group Sikhs for Justice (SFJ) in India.[311] SFJ alleged that Facebook took such action without explanation, "'on its own or on the behest of the Government of India,' because of discrimination against Plaintiff and Plaintiff's members on the grounds of race, religion, ancestry, and national origin."[312] Even when SFJ requested an explanation, Facebook allegedly didn't respond.[313] The district court for the Northern District of California ruled that Section 230(c)(1) provided immunity for Facebook's decisions,[314] but the court's ruling is based on the misreading of (c)(1) discussed above. Section 230(c)(2) is the correct provision that governs content moderation. Under (c)(2), SFJ's claims would not be barred, assuming the complaint satisfies the general requirement of pleading sufficient allegations to support the claims.[315] SFJ claimed that Facebook completely blocked SFJ's page based not on the content, but simply on the *religious identity* of SFJ's members who "oppos[ed] the forced conversions of religious minorities to Hinduism that have allegedly taken place in India since the election of Prime Minister Narendra Modi."[316] I agree with Justice Thomas's suggestion in *Malwarebytes* that this case may have been wrongly decided.[317] Assuming there were sufficient allegations of discriminatory animus in the complaint, the alleged conduct would fall outside of Section 230(c)(2) immunity. It is noteworthy that Facebook's content moderation decisions in India have also been questioned as showing favoritism to the ruling party instead of being based on Facebook's

---

Senate Committee in 2020 is considered "[p]ost-enactment legislative history" and, at least to some judges, "not a legitimate tool of statutory interpretation." Bruesewitz v. Wyeth LLC, 562 U.S. 223, 242 (2011); *see PACT Act Hearings*, *supra* note 139, at 17 (testimony of Chris Cox, Former Member, U.S. House of Representatives).

311. *See* Sikhs for Justice "SFJ," Inc. v. Facebook, Inc., 144 F. Supp. 3d 1088, 1090 (N.D. Cal. 2015), *aff'd*, Sikhs for Justice, Inc. v. Facebook, Inc., 697 F. App'x 526 (9th Cir. 2017).

312. *Id.* at 1090.

313. *Id.*

314. *Id.* at 1096.

315. *See generally* Ashcroft v. Iqbal, 556 U.S. 662, 678–79 (2009) ("But where the well-pleaded facts do not permit the court to infer more than the mere possibility of misconduct, the complaint has alleged—but it has not 'show[n]'—'that the pleader is entitled to relief.'" (quoting Fed. R. Civ. P. 8(a)(2)) (citing Bell Atl. Corp. v. Twombly, 550 U.S. 544 (2007))).

316. *See Sikhs for Justice "SFJ," Inc.*, 144 F. Supp. 3d at 1090.

317. *See* Malwarebytes, Inc. v. Enigma Software Grp. USA, LLC, 141 S. Ct. 13, 18 (2020).

own stated policies.[318] If Facebook removes content because of the user's political affiliation, even or especially if at the behest of the ruling political party, such a decision does not fall within Section 230(c)(2). It is a decision based on the identity of the user, not the actual material.

d. A decision following a company's own stated policies can be evidence of good faith

The more difficult question involves a situation in which an internet platform says it removed user content that violates a stated policy of the company (typically its community guidelines) and explains the reason to the user, but the user contends the reason was pretextual and based instead on a discriminatory reason, such as the user's political party.

Domen takes the right approach. An internet platform's removal of user content that violates the platform's stated policies regarding "objectionable" content falls within Section 230(c)(2), absent specific allegations of "bad faith" to satisfy the standards of pleading (and eventual proof at trial).[319] We might call this approach the "stated policy" approach to "good faith" moderation—similar to Cox's interpretation described above.[320] As long as the internet platform has a stated policy addressing such moderation of content, the platform's decision to enforce its stated policy would fall within good faith moderation—absent some evidence that the platform's decision was pretextual and was based instead on an improper reason unrelated to the content itself.

This approach allows internet platforms to engage in viewpoint discrimination, but gives them incentives to provide adequate notice to their users of what content is impermissible to share. For example, in 2019, Facebook expanded its ban on content promoting white supremacy to include white nationalism and white separatism on

318. *See, e.g.*, Newley Purnell & Jeff Horwitz, *Facebook Executive Supported India's Modi, Disparaged Opposition in Internal Messages*, WALL ST. J. (Aug. 30, 2020, 1:42 PM), https://www.wsj.com/articles/facebook-executive-supported-indias-modi-disparaged-opposition-in-internal-messages-11598809348; Newley Purnell, *Facebook's Top Public Policy Executive in India Steps down*, WALL ST. J. (Oct. 27, 2020, 12:24 PM), https://www.wsj.com/articles/facebook-s-top-public-policy-executive-in-india-steps-down-11603807845.

319. *See* Domen v. Vimeo, Inc., 433 F. Supp. 3d 592, 603–04 (S.D.N.Y. 2020).

320. *See supra* note 141 and accompanying text.

Facebook.[321] Under the "stated policy" approach, Facebook acts in "good faith" by moderating such content that it considers "objectionable," according to its stated policy, even though it constitutes overt viewpoint discrimination.[322] Conversely, imagine that a new internet platform marketed itself as the "social media for conservatives" and adopted a community standard banning as "objectionable" content promoting Antifa, immigration, or radical left-wing ideas. Under the "stated policy" approach, content moderation enforcing this stated policy would be in "good faith," even though discriminating based on political viewpoint. The lack of a stated policy for a company's decision to remove user content does not necessarily indicate "bad faith," but the internet platform still must explain why it removed the user content with a reason that is not arbitrary.[323]

This approach is consistent with the Section 230(c)(2) case law. For example, a district court in Florida rejected a motion to dismiss by Google and recognized sufficient pleading of bad faith based on the allegations that Google had delisted 231 websites affiliated with the plaintiff's search engine optimization company from Google's search engine "solely based upon the websites' affiliation with e-ventures, which did not fall within any of Google's listed reasons that it would remove a website from its search results."[324] However, courts have rejected arguments to defeat Section 230(c)(2) immunity simply based on the internet platform's "failure to remove all such [violating] content" from its site, which is somewhat akin to a selective enforcement claim.[325] Courts have also been deferential to the internet platforms' "good faith" determination of "objectionable" material.[326] In deciding motions to dismiss, courts have tended to be

---

321. *Standing Against Hate*, FACEBOOK (Mar. 27, 2019), https://about.fb.com/news/2019/03/standing-against-hate [https://perma.cc/YSF7-KQ9C]

322. *See generally PACT Act Hearings*, *supra* note 139, at 17 (testimony of Chris Cox, Former Member, U.S. House of Representatives).

323. *Id.*

324. e-ventures Worldwide, LLC v. Google, Inc., 188 F. Supp. 3d 1265, 1269–71, 1273 (M.D. Fla. 2016). One court held that an internet company's failure to "respond to Plaintiff's repeated requests for an explanation why it continually blocked Plaintiff's outgoing e-mail" might be bad faith. *See* Smith v. Trusted Universal Standards in Elec. Transactions, Inc., No. 09-4567, 2011 WL 900096, at *9 (D.N.J. Mar. 15, 2011); Eric Goldman, *Online User Account Termination and 47 U.S.C. § 230(c)(2)*, 2 U.C. IRVINE L. REV. 659, 665 (2012).

325. *See* Pennie v. Twitter, Inc., 281 F. Supp. 3d 874, 890 (N.D. Cal. 2017).

326. *See, e.g.*, Holomaxx Techs. v. Microsoft Corp., 783 F. Supp. 2d 1097, 1104 (N.D. Cal. 2011) (finding that "it is clear from the allegations of the complaint itself

skeptical of "lack of good faith" claims asserted against internet companies with conclusory allegations that do not satisfy the requirement of adequate pleading.[327]

The "stated policy" approach doesn't give internet platforms complete discretion to adopt any content moderation policy to fall within Section 230's immunity. Section 230 is limited to moderation of "material," not users.[328] Content moderation must be based on the actual content posted by the user, as the text of Section 230 indicates ("material that the provider or user considers to be . . . otherwise objectionable").[329] Thus, even if an internet platform had a stated policy of promoting content consistent with a certain religion or political party, the platform's decision to remove "objectionable" content solely based on the person's political affiliation would not qualify as action in "good faith." Internet platforms can suspend or take actions against the accounts of users for repeated or other violations of their policies, such as terrorist groups,[330] but Section 230(c)(2)'s immunity applies only to moderation of "material."

The lines between political affiliation and political viewpoint may get blurred. For example, QAnon is a conspiracy theory about the so-called "deep state."[331] It is not a political party per se, although believers in QAnon are often portrayed as right-wing supporters of Trump.[332] In July 2020, Twitter suspended 7,000 accounts involving "'QAnon' activity" for violations of its policy against content

---

that Microsoft reasonably could conclude that Holomaxx's e-mails were 'harassing' and thus 'otherwise objectionable'").

327. *See, e.g.*, Domen v. Vimeo, Inc., 433 F. Supp. 3d 592, 603–04 (S.D.N.Y. 2020) (rejecting lack of good faith allegation where complaint merely alleged "that Vimeo applied its Guidelines to remove Plaintiffs' videos, since such videos violated the Guidelines"); e360Insight, LLC v. Comcast Corp., 546 F. Supp. 2d 605, 609 (N.D. Ill. 2008) (finding inadequate pleading of lack of good faith); Donato v. Moldow, 865 A.2d 711, 727 (N.J. Super. Ct. App. Div. 2005) (rejecting the argument of bad faith based on the bare allegation that an operator of a community bulletin board knew the plaintiffs and "published the defamatory statements with actual malice").

328. 47 U.S.C. § 230.

329. *Id.* § 230(c)(2)(A).

330. *See About*, GLOBAL INTERNET F. TO COUNTER TERRORISM, https://www.gifct.org/about [https://perma.cc/B3KL-VDEY].

331. *See* Adrienne LaFrance, *The Prophecies of Q*, ATLANTIC (June 2020), https://www.theatlantic.com/magazine/archive/2020/06/qanon-nothing-can-stop-what-is-coming/610567/?utm_source=newsletter&utm_medium=email&utm_campaign=atlantic-daily-newsletter&utm_content=20200722&silverid-ref=NTg4ODEzMDIzNTA5S0.

332. *Id.*

potentially leading to offline harm and against coordinated use of multiple accounts.[333] Twitter stated that it would "[n]o longer serve content and accounts associated with QAnon in Trends and recommendations" and would "[b]lock URLs associated with QAnon from being shared on Twitter."[334] Twitter's QAnon decision was controversial.[335] Does Twitter's moderation of QAnon qualify as "good faith" moderation of "objectionable" material? It is difficult to determine without access to the actual content posted by each user. If the content violated Twitter's stated policies, Twitter's removal of the objectionable material falls within Section 230(c)(2). But the decision to terminate a user account for such repeated violations would fall within the terms of service agreement Twitter has with its users, not Section 230(c)(2).

5. Lack of immunity does not establish liability

Before turning to the debate over alleged political bias among internet platforms, it is important to recognize that an internet company's "bad faith" removal of third-party content that is not immunized under Section 230(c)(2) does not necessarily spell liability. The third party who created and posted the content still must raise a cause of action. What might that be?

Zipursky suggested one possibility: a voluntary undertaking tort claim, which Section 230(c) itself suggests by the reference in the title to "Good Samaritan," who voluntarily takes on a duty to assist where none existed.[336] The undertaking presumably would be moderating content according to the company's community standards, and the company would have to exercise reasonable care (i.e., a negligence

333. *See* Rachel Lerman & Elizabeth Dwoskin, *Twitter Crackdown on Conspiracy Theories Could Set Agenda for Other Social Media*, WASH. POST (July 22, 2020, 5:49 PM), https://www.washingtonpost.com/technology/2020/07/22/twitter-bans-qanon-accounts.

334. *See* Twitter Safety (@TwitterSafety), TWITTER (July 21, 2020, 8:00 PM), https://twitter.com/TwitterSafety/status/1285726277719199746.

335. *See, e.g.*, Evelyn Douek, *Twitter Brings down the Banhammer on QAnon*, LAWFARE (July 24, 2020, 2:56 PM), https://www.lawfareblog.com/twitter-brings-down-banhammer-qanon [https://perma.cc/355L-DZB3] (discussing whether Twitter's policy was less about the content of QAnon tweets and more about Twitter's need for a "good news cycle"); Abby Ohlheiser, *It's Too Late to Stop QAnon with Fact Checks and Account Bans*, MIT TECH. REV. (July 26, 2020), https://www.technologyreview.com/2020/07/26/1005609/qanon-facebook-twitter-youtuube (arguing Twitter's new policy does not limit QAnon as much as it seems).

336. *See* Zipursky, *supra* note 158, at 31–34.

standard) in performing that duty.[337] Because this is not a strict liability standard, such an inquiry would afford an internet platform some leeway to make even mistaken removal decisions, given the sheer scale of third-party content that must be reviewed for content moderation. For example, an automated removal of content through a platform's artificial intelligence might satisfy the standard of reasonable care, even if it resulted in mistaken takedowns, perhaps if the decisions could later be corrected by user appeals as is common among internet platforms.

Another possibility might be a breach of contract claim if the internet platform did not even follow its own terms of service. Although such contractual claims are preempted when Section 230(c)(2) immunity applies,[338] presumably such claims are not preempted when the immunity is inapplicable (e.g., bad faith moderation). This approach may make internet platforms skittish about "over-promising" in their terms of service or community standards regarding content moderation—which might have the unintended consequence of encouraging internet platforms to be even less transparent in their terms of service with users. On the other hand, if the internet platforms have a stated policy for content moderation and simply follow it, their decisions adhering to the stated policy would fall within Section 230(c)(2) immunity.

As noted earlier, the dormant Commerce Clause would bar state law claims against an internet service that resulted in excessive burden on interstate commerce. Even if a cause of action attaches liability to "bad faith" content moderation in one state, such liability might violate the dormant Commerce Clause by burdening commerce in other states that do not recognize such liability.[339] For example, the Second Circuit held that a Vermont statute prohibiting the transfer to minors of any sexually explicit material that was "harmful to minors" violated the dormant Commerce Clause.[340] The court even suggested that the internet may "fall[] within the class of subjects that are protected from State regulation because they

337. *See generally* Frye v. Medicare-Glaser Corp., 605 N.E.2d 557, 560–61 (Ill. 1992).

338. *See* Batzel v. Smith, 333 F.3d 1018, 1030 n.14 (9th Cir. 2003); *cf.* Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1106–07 (9th Cir. 2009) (for Section 230(c)(1), recognizing that breach of contract claims, including promissory estoppel, are not barred because they are based on promising, not publishing).

339. *See* Am. Booksellers Found. v. Dean, 342 F.3d 96, 103 (2d Cir. 2003).

340. *Id.* at 103–04.

'imperatively demand[] a single uniform rule.'"[341] Imagine that Texas recognized a claim against an internet platform for "bad faith" moderation of third-party content based on a bias against the user's political affiliation, whereas California allowed such moderation. Although it might be technologically feasible for an internet platform to try to differentiate its website based on which state it is viewed, as some websites do by country, it could create an excessive burden on interstate commerce by subjecting internet companies to a patchwork of state laws and a need to create a balkanized internet for each state, depending on what type of content moderation decisions gave rise to liability.[342]

D. Accusations of Political Bias and Proposed Amendments to Section 230 to Require Political Neutrality

The prior section established that a proper reading of Section 230 recognizes a potential disqualification from Section 230(c)(2) for a "bad faith" removal of content based solely on the political affiliation of the third-party. Because courts have yet to decide the issue and because alleged "anti-conservative" bias has become a hot-button issue, it is not surprising that Republican lawmakers have proposed several bills to amend Section 230 to require, expressly or in effect, some form of political viewpoint neutrality as a prerequisite for internet platforms to qualify for Section 230 immunity.[343] This Section provides a summary of these bills, as well as Trump's Executive Order and the DOJ's recommendations on Section 230 to Congress. It is unclear which bill, if any, has a realistic chance of enactment by Congress. But it appears likely that the fervor over amending or

---

341. *Id.* at 104 (second alteration in original) (quoting Cooley v. Bd. of Wardens, 53 U.S. (12 How.) 299, 319 (1852)). *But see* Jack L. Goldsmith & Alan O. Sykes, 110 YALE L.J. 785, 786–87 (2001) (criticizing as incorrect the Second Circuit's suggestion).

342. *Am. Booksellers Found.*, 342 F.3d at 103 ("Because the internet does not recognize geographic boundaries, it is difficult, if not impossible, for a state to regulate internet activities without 'project[ing] its legislation into other States.'" (quoting Healy v. Beer Inst., 491 U.S. 324, 334 (1989))).

343. Congress is considering other amendments to Section 230 (e.g., EARN IT Act, BADS ADS Act) that do not require some form of political neutrality or eliminate the catchall phrase "otherwise objectionable" in Section 230(c)(2). *See, e.g.*, Eliminating Abusive and Rampant Neglect of Interactive Technologies Act of 2020, S. 3398, 116th Cong. (2020); Behavioral Advertising Decisions Are Downgrading Services Act, S. 4337, 116th Cong. (2020). These bills are not discussed.

outright repealing Section 230 will lead to some action in Congress. The following summary of the bills shows the dramatic changes to Section 230 being considered, including greatly limiting the discretion internet platforms have in what content can be removed, requiring greater disclosure to users of the platforms' content moderation policies, imposing a viewpoint neutrality requirement, authorizing FTC oversight over internet platforms' content moderation, defining the "good faith" requirement or "bad faith," and recognizing a private right of action against internet platforms for content moderation that violates a requirement imposed on internet platforms. And a complete repeal of Section 230 is on the table.[344]

1. Limiting Section 230 Immunity to Good Samaritans Act: proscribing intentional selective enforcement of content moderation policy

The Limiting Section 230 Immunity to Good Samaritans Act,[345] introduced by Senator Josh Hawley (R-MO) in June 2020, would amend Section 230(c)(1) immunity by adding a condition for a newly recognized class of large "edge providers" who must write terms of service that "describe any policies of the edge provider relating to restricting access to or availability of material" and promise they "design and operate the provided service in good faith."[346] Good faith is defined as "the provider acts with an honest belief and purpose, observes fair dealing standards, and acts without fraudulent intent."[347] The bill defines lack of good faith to include "the intentionally selective enforcement of the terms of service of the interactive computer service, including the *intentionally selective enforcement of policies of the provider relating to restricting access to or availability of material*," meaning content moderation.[348] Bad faith intentional selective enforcement also applies to algorithmic decisions "if the provider knows, or acts in reckless disregard of the fact, that the algorithm selectively enforces those terms."[349] The bill creates a cause of action for users to sue internet edge providers for intentional selective enforcement and to recover $5,000 in statutory damages or

344. It goes beyond the scope of this Article to provide a full critique of each proposal.
345. S. 3983, 116th Cong. (2020).
346. *Id.* § 2.
347. *Id.*
348. *Id.* (emphasis added).
349. *Id.*

actual damages.[350] Edge providers are large interactive computer services that have more than thirty million users in the United States or more than 300 million users worldwide, plus more than $1.5 billion in annual global revenue, but excluding 501(c)(3) nonprofits.[351] Notably, the intentional selective enforcement claim under this bill applies to *all* content moderated by edge providers, not just content posted by political candidates or in political ads.

2. Ending Support for Internet Censorship Act (ESICA): requiring FTC immunity certification of politically unbiased moderation

Senator Hawley introduced an earlier bill in June 2019 titled Ending Support for Internet Censorship Act.[352] This bill would add a "[r]equirement of politically unbiased content moderation by covered companies" as a prerequisite to immunity under either Section 230(c)(1) or 230(c)(2).[353] Covered companies are interactive computer services, other than 501(c)(3) nonprofits, that in the last twelve-month period had more than thirty million active monthly users in the United States, more than 300 million active monthly users worldwide, or more than $500 million in global revenues.[354] Covered companies are required to obtain "an immunity certification from the Federal Trade Commission" (FTC) lasting for two years by proving, by clear and convincing evidence, "that the provider does not (and, during the 2-year period preceding the date on which the provider submits the application for certification, did not) moderate information provided by other information content providers in a *politically biased manner*."[355] Politically biased moderation is defined as:

> (I) the provider moderates information provided by other information content providers in a manner that—
>
> (aa) is designed to negatively affect a political party, political candidate, or political viewpoint; or
>
> (bb) disproportionately restricts or promotes access to, or the availability of, information from a political party, political candidate, or political viewpoint; or

350. *Id.*

351. *Id.* Representative Tedd Budd introduced the same bill in the House. H.R. 8596, 116th Cong. (2020).

352. S. 1914, 116th Cong. (2019).

353. *Id.* § 2.

354. *Id.*

355. *Id.* (emphasis added).

> (II) an officer or employee of the provider makes a decision about moderating information provided by other information content providers that is motivated by an intent to negatively affect a political party, political candidate, or political viewpoint.[356]

This broad definition appears to recognize both intentional discrimination and disparate impact claims even without discriminatory intent of the provider. The bill recognizes a limited exception for business necessity and for moderation decisions of employees "if the provider, immediately upon learning of the actions of the employee—(aa) publicly discloses in a conspicuous manner that an employee of the provider acted in a politically biased manner with respect to moderating information content; and (bb) terminates or otherwise disciplines the employee."[357] During the certification process, the FTC must establish a process for public participation, including public submissions and attendance at hearings, and the FTC must vote by one more than majority vote (meaning at least four members) for a certification to be approved.[358] In sum, this bill requires an elaborate FTC certification process and a very high burden for internet platforms to prove, by clear and convincing evidence, that their content moderation is not politically biased.

3. Stop the Censorship Act (SCA): limiting Section 230(c)(2) to content moderation of unlawful material

The Stop the Censorship Act,[359] introduced by Representative Paul Gosar (R-AZ-4), would repeal the phrase "material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected" in Section 230(c)(2).[360] The bill would then limit immunity to a company's moderation of "unlawful material,"[361] plus "any action taken to provide users with the option to restrict access to any other material, whether or not such material is constitutionally protected."[362] The effect of the bill apparently would be that internet platforms have no immunity for their content moderation of *any* lawful material, including nudity,

356. *Id.*
357. *Id.*
358. *Id.*
359. H.R. 4027, 116th Cong. (2020).
360. *Id.* § 2.
361. *Id.*
362. *Id.*

pornography, sexually explicit material, hate speech, misinformation, or harassment that is not unlawful. But it would extend immunity to the platform's actions in providing their users with the ability to restrict access to such materials.[363] Presumably, this immunity would cover a set of filtering options provided for users to screen out sexually explicit material, hate speech, racist speech, conspiracy theories, and misinformation. At least indirectly, SCA would treat viewpoint discrimination by an internet platform as falling outside of Section 230(c)(2) except for moderation involving unlawful material.

4. Stopping Big Tech's Censorship Act (SBTCA): requiring content moderation to follow First Amendment-style restrictions, including viewpoint neutrality

The Stopping Big Tech Censorship Act,[364] introduced by then-Senator Kelly Loeffler (R-GA), would require internet platforms to satisfy new prerequisites for Section 230 immunity.[365] First, the bill would require internet platforms to "take[] reasonable steps to prevent or address the unlawful use of the interactive computer service or unlawful publication of information on the interactive computer service," to qualify for the immunity from defamation and other claims based on the content of their users.[366] Second, the bill would allow internet platforms to engage in content moderation of lawful speech—what the bill calls "constitutionally protected material"—only under the following First Amendment-style conditions drawn from various Supreme Court jurisprudence that apply to government restrictions of speech: "(I) the action is taken in a *viewpoint-neutral manner*; (II) the restriction limits only the time, place, or manner in which the material is available; and (III) there is a compelling reason for restricting that access or availability."[367] As the bill indicates, it subjects internet platforms to these conditions "regardless of whether the right is otherwise enforceable against a nongovernmental entity,"[368] presumably indicating the lack of a state action requirement for internet platforms.

---

363. *Id.*
364. S. 4062, 116th Cong. (2020).
365. *See id.* § 2 (detailing some of the new prerequisites for immunity, including how content is restricted and notice requirements).
366. *Id.*
367. *Id.* (emphasis added).
368. *Id.*

5. Stop Suppressing Speech Act: deleting "otherwise objectionable" from Section 230(c)(2)

Loeffler proposed a second bill titled Stop Suppressing Speech Act.[369] Similar to Gosar's Stop the Censorship Act, Loeffler's bill eliminates the catchall "otherwise objectionable" material in Section 230(c)(2) and would limit the permissible bases for content moderation to the following:

> any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, or harassing, *that the provider or user determines to be unlawful, or that promotes violence or terrorism,* whether or not such material is constitutionally protected.[370]

6. Platform Accountability and Consumer Transparency Act (PACT Act): requiring internet platforms to publish "an acceptable use policy" and provide live customer service for content moderation

One bipartisan-sponsored bill to reform Section 230 is the Platform Accountability and Consumer Transparency Act (PACT Act), sponsored by Senators Brian Schatz (D-HI) and John Thune (R-SD).[371] The PACT Act would require internet platforms to "publish an acceptable use policy . . . in a location that is easily accessible to the user."[372] In addition, the acceptable use policy must "reasonably inform users about the types of content that are allowed" and "explain the steps the provider takes to ensure content complies with the acceptable use policy."[373] The acceptable use policy must also "explain the means by which users can notify the provider of potentially policy-violating content, illegal content, or illegal activity, which shall include" a complaint system and "a live company representative to take user complaints through a toll-free telephone number."[374] The PACT Act sets forth the requirements and decision deadlines for content moderation in great detail.[375] The FTC is given authority to enforce these requirements and treat violations as unfair or deceptive practices under the Federal Trade Commission Act.[376]

---

369. Stop Suppressing Speech Act of 2020, S. 4828, 116th Cong.
370. *Id.* (emphasis added).
371. Platform Accountability and Consumer Transparency Act, S. 4066, 116th Cong. (2020).
372. *Id.* § 5.
373. *Id.*
374. *Id.*
375. *See id.* (outlining the requirements for the acceptable use policy).
376. 15 U.S.C. §§ 41–58; S. 4066 § 5.

Furthermore, the PACT Act would add an additional requirement to Section 230(c) that would disqualify internet platforms from Section 230 immunities if the platforms had "knowledge of the illegal content or illegal activity" on their platforms, but failed to remove it "within 24 hours of acquiring that knowledge, subject to reasonable exceptions based on concerns about the legitimacy of the notice."[377] The PACT Act does not require political neutrality or address whether an acceptable use policy can moderate content based on a political viewpoint that the platform deems objectionable.

7. The Online Freedom and Viewpoint Diversity Act (OFVDA): limiting Section 230(c)(2) to removal of "obscene, lewd, lascivious, filthy, excessively violent, harassing, promoting self-harm, promoting terrorism, or unlawful" material

The Online Freedom and Viewpoint Diversity Act[378] (OFVDA), introduced on September 8, 2020, is another Republican bill intended to reform Section 230.[379] It was introduced by Senator Roger Wicker (R-MS), Chairman of the Senate Committee on Commerce, Science, and Transportation, and co-sponsored by Senators Lindsey Graham (R-SC), Chairman of the Senate Committee on the Judiciary, and Marsha Blackburn (R-TN).[380] This bill is similar in approach to Senator Gosar's proposed Stop the Censorship Act. OFDVA would change "otherwise objectionable" material to material "promoting self-harm, promoting terrorism, or unlawful."[381] The bill also would change Section 230(c)(1)'s subjective standard (i.e., what the internet service "considers to be" objectionable content) to require instead that the internet service "ha[ve] an objectively reasonable belief" the content in question falls within a listed category of removable content.[382] The bill adds a provision to clarify that Section 230(c)(1), the first type of immunity, does not apply to content moderation, which is covered by the second type of immunity in Section 230(c)(2).[383] As I have explained above, this bifurcation—or two different immunities—already exists under the current Section

377. *Id.* § 6.
378. S. 4534, 116th Cong. (2020).
379. *Id.*
380. *Id.*
381. *Id.* § 2.
382. *Id.*
383. *Id.*

230, notwithstanding some courts' misreading.[384] The bill amends the current definition of "information content provider," who is treated as the publisher or speaker of content under Section 230(c)(1)—typically the user who has posted the material.[385] The bill would expand the definition to include when "a person or entity editorializes or affirmatively and substantively modifies the content of another person or entity."[386] Such editing or modifying would make that person also an information content provider. It is unclear whether this provision is intended to reach instances in which internet platforms add a label or notation to a user's post that it violates the platform's community standard. Sen. Hawley also proposed an amendment to OFDVA to include a private right of action to sue large internet platforms ("edge providers") for moderation of user material "that is not taken in good faith."[387] However, the Senate Judiciary Committee voted against the proposed Hawley amendment.[388] By shrinking the permissible bases for content moderation under Section 230(c)(2) without the catchall for "otherwise objectionable" material, OFDVA would greatly limit the discretion internet platforms have in content moderation. Moderation of any content that does not fall within one of the categories would not receive Section 230 immunity.

8. Sunset for and repeal of Section 230 proposed by Senator Graham

On December 15, 2020, Senator Graham proposed yet another bill to reform Section 230.[389] The bill would create a sunset of January 1, 2023 at which time Section 230 would expire unless Congress enacts a new law for the section during the intervening two-year period.[390] It is the first bill to propose a repeal of Section 230, although it is framed

384. *See* 47 U.S.C. § 230(c)(2)(A); *supra* note 148 and accompanying text (explaining the bifurcation of Section 230 immunity).

385. *See* Online Freedom and Viewpoint Diversity Act, S. 4534 § 2 (listing the changes to the definition of information content provider the bill proposes).

386. *Id.*

387. *See* S. 4632, 116th Cong. (2020).

388. *See* Steven Overly, *Congress Aims to Avert Shutdown*, POLITICO (Dec. 11, 2020, 10:00 AM), https://www.politico.com/newsletters/morning-tech/2020/12/11/congress-aims-to-avert-shutdown-792228 [https://perma.cc/D5KV-UU5T].

389. *See* S. 5020, 116th Cong. (2020); Tal Axelrod, *Graham Introduces Bill to Repeal Tech Liability Shield Targeted by Trump*, HILL (Dec. 15, 2020, 5:40 PM), https://thehill.com/homenews/senate/530364-graham-introduces-bill-to-repeal-tech-liability-shield-by-2023 [https://perma.cc/7AER-BC4Q].

390. *Id.*

with the intention of coming up with an alternative approach. (In a political maneuver in response to Trump's threat to veto an omnibus budget bill that included COVID-relief payments of $600, Senator Mitch McConnell proposed a bill that tied increasing payments to $2,000 to a repeal of Section 230; the bill failed, however.[391])

9. President Trump's executive order and FCC's proposed Section 230 rulemaking

Trump's Executive Order on Preventing Online Censorship purports to "clarify" immunity under Section 230.[392] According to the Order,

> the immunity should not extend beyond its text and purpose to provide protection for those who purport to provide users a forum for free and open speech, but in reality use their power over a vital means of communication to engage in deceptive or pretextual actions stifling free and open debate by censoring certain viewpoints.[393]

For the immunity under Section 230(c)(2), the Order states:

> It is the policy of the United States to ensure that, to the maximum extent permissible under the law, this provision is not distorted to provide liability protection for online platforms that—far from acting in "good faith" to remove objectionable content—instead engage in deceptive or pretextual actions (often contrary to their stated terms of service) to stifle viewpoints with which they disagree.[394]

The Order takes the view that an internet platform would not qualify for immunity if it engaged in the "stif[ling] [of] viewpoints."[395]

An Executive Order cannot alter or amend an act of Congress.[396] Some may question the legal authority for the actions directed by Trump's Executive Order. The Order invokes "the Constitution and the laws of the United States of America" as authority, without greater specificity.[397]

---

391. *See* Makena Kelly, *McConnell Ties Full Repeal of Section 230 to Push for $2,000 Stimulus Checks*, VERGE (Dec. 29, 2020, 5:30 PM), https://www.theverge.com/2020/12/29/22204976/section-230-senate-deal-stimulus-talks-checks.

392. Exec. Order No. 13,925 § 2, 85 Fed. Reg. 34,079, 34,080 (May 28, 2020).

393. *Id.* § 2(a).

394. *Id.*

395. *Id.*

396. *See* Chamber of Com. of the U.S. v. Reich, 74 F.3d 1322, 1339 (D.C. Cir. 1996) (holding that an Executive Order that conflicted with the National Labor Relations Act was invalid).

397. Exec. Order No. 13,925 § 1, 85 Fed. Reg. at 34,079.

The Order instructs "the Secretary of Commerce (Secretary), in consultation with the Attorney General, and acting through the National Telecommunications and Information Administration (NTIA), [to] file a petition for rulemaking with the Federal Communications Commission (FCC) requesting that the FCC expeditiously propose regulations to clarify" Section 230.[398] Section 230 does not contain a delegation of rulemaking authority to the FCC, but the section was contained in Title V of the Telecommunications Act of 1996,[399] which amended the Communications Act of 1934,[400] a statute that includes a provision authorizing the FCC to conduct rulemaking for its regulation of common carriers and "the provisions of this chapter."[401] FCC Commissioner Michael O'Rielly, a Republican appointee, expressed "deep reservations" that rulemaking authority under the Communications Act of 1934 extended to the later-added Section 230; afterwards, Trump withdrew his nomination of O'Rielly for another term as FCC Commissioner.[402]

The Executive Order seeks to have the FCC issue regulations interpreting Section 230 so that, effectively, the "good faith" requirement of Section 230(c)(2) also applies to subsection (c)(1).[403] Also, the Order seeks regulations to define "good faith" to exclude content moderation decisions that are "(A) deceptive, pretextual, or inconsistent with a provider's terms of service; or (B) taken after failing to provide adequate notice, reasoned explanation, or a meaningful opportunity to be heard."[404] The Order instructs the Secretary of Commerce, "acting through the [NTIA], [to] file a petition for rulemaking with the [FCC] requesting that the FCC expeditiously propose regulations to clarify" issues related to:

> (i) the interaction between subparagraphs (c)(1) and (c)(2) of [S]ection 230, in particular to clarify and determine the circumstances under which a provider of an interactive computer service that

398. *Id.* § 2(b), 85 Fed. Reg. at 34,081.

399. Pub. L. 104-104, 110 Stat. 56 (1996).

400. 47 U.S.C. §§ 151–621.

401. 47 U.S.C. § 201(b) ("The Commission may prescribe such rules and regulations as may be necessary in the public interest to carry out the provisions of this chapter.").

402. *See* Russell Brandom, *President Trump Withdraws FCC Renomination After 5G Controversy*, VERGE (Aug. 3, 2020, 6:01 PM), https://www.theverge.com/2020/8/3/21353233/orielly-fcc-nomination-withdrawn-trump-ligado-5g-230.

403. *See* Exec. Order No. 13,925 § 2, 85 Fed. Reg. at 34,080–81.

404. *Id.* § 2(b)(ii), 85 Fed. Reg. at 34,081.

> restricts access to content in a manner not specifically protected by subparagraph (c)(2)(A) may also not be able to claim protection under subparagraph (c)(1)[] . . . [and] (ii) the conditions under which an action restricting access to or availability of material is not "taken in good faith" within the meaning of subparagraph (c)(2)(A) of [S]ection 230.[405]

On July 27, 2020, invoking the delegation of rulemaking authority under the Communications Act of 1934, the NTIA filed its petition asking the FCC to conduct a rulemaking on Section 230 along the lines indicated in Trump's Executive Order.[406] FCC Chairman Pai later stated his intent "to move forward with a rulemaking to clarity [Section 230's] meaning."[407] However, after Trump lost the 2020 election, Pai said there was not sufficient time to do so because he was stepping down on January 20, 2021.[408]

The Executive Order instructs another agency, the FTC, to investigate large online platforms, such as Twitter and Facebook, for unfair and deceptive acts related to content moderation.[409] The Order instructs the Attorney General to establish a "working group regarding the potential enforcement of State statutes that prohibit online platforms from engaging in unfair or deceptive acts or practices."[410]

10. Department of Justice recommendations on Section 230

In June 2020, the DOJ issued a report on "Section 230—Nurturing Innovation or Fostering Unaccountability."[411] DOJ recommended four types of reforms for Congress to consider: (1) "incentivizing online platforms to address illicit content"; (2) "making [it] clear that

---

405. *Id.* § 2(b).

406. National Telecommunications and Information Administration, Petition for Rulemaking in the Matter of Section 230 of the Communications Act of 1934 (July 27, 2020), https://www.ntia.gov/files/ntia/publications/ntia_petition_for_rulemaking_7.27.20.pdf [https://perma.cc/M78T-8C7Y], at 1.

407. *Statement of Chairman Pai on Section 230*, FCC (Oct. 15, 2020), https://docs.fcc.gov/public/attachments/DOC-367567A1.pdf [https://perma.cc/V9PX-7MN6].

408. *See* Todd Shields & Ben Brody, *FCC Chair Punts Social Media Regulation Trump Sought to Congress*, BLOOMBERG (Jan. 8, 2021, 1:01 PM), https://www.bloomberg.com/news/articles/2021-01-08/fcc-chair-says-he-s-dropping-social-media-order-trump-demanded.

409. Exec. Order No. 13,925 § 4(c), 85 Fed. Reg. at 34,082.

410. *Id.* § 5(a).

411. U.S. Dep't of Just., Section 230—Nurturing Innovation or Fostering Unaccountability? (2020).

the immunity provided by Section 230 does not apply to civil enforcement [actions brought] by the federal government"; (3) "clarify[ing] that federal antitrust claims are not covered by Section 230 immunity"; and (4) "clarify[ing] the text and original purpose of the statute in order to promote free and open discourse online and encourage greater transparency between platforms and users."[412]

For the first type of reform, DOJ recommended a "Bad Samaritan carve-out" in Section 230 for "online platforms that purposefully promote[], solicit[], or facilitate[] criminal activity by third parties."[413] Similar to the Gosar and Wicker bills discussed above, DOJ recommended that Congress remove the catchall for "otherwise objectionable" material and limit it to "material the platform believes, in good faith, violates federal law or promotes violence or terrorism."[414] DOJ took the view that internet platforms can still moderate content beyond the proposed amended categories as long as they are consistent with their terms of service.[415] DOJ states: "Online platforms are often protected by their terms of service when removing content that violates the platform's rules, whether or not that content falls into the categories of (c)(2)."[416] Yet, unlike other sections in the report, DOJ cites no case law or legal authorities to support this assertion.

Similar to the Executive Order, DOJ recommended that Congress define "good faith" content moderation in Section 230(c)(2) with four principles geared around (1) the disclosure of the content-moderation practices in the company's terms of service, (2) content moderation that is consistent with the terms of service "and with any official representations regarding the platform's content-moderation policies," (3) the company having an "objectively reasonable belief" the content falls with Section 230(c)(2), and (4) "the platform [ ] supply[ing] the provider of the content with a timely notice explaining with particularity the factual basis for the restriction of access, unless the provider reasonably believes that the content relates to criminal activity or notice would risk imminent harm to others."[417] Finally, DOJ proposed that Congress continue to reject the moderator's dilemma created by *Stratton Oakmont* by "adding a

412. *Id.* at 3–4.
413. *Id.* at 3, 14.
414. *Id.* at 21.
415. *Id.*
416. *Id.*
417. *Id.* at 22.

provision to make clear that a platform's decision to moderate content either under (c)(2) or consistent with its terms of service does not automatically render it a publisher or speaker for all other content on its service."[418]

The Trump Administration's planned actions regarding Section 230 were preempted with the election of Joe Biden as the next President. However, Biden himself is no fan of Section 230 and has called for its repeal for a reason different from Trump's: Section 230 allows Facebook to "propagat[e] falsehoods they know to be false."[419] Thus, even under Biden's Administration, Section 230 reform seems likely.

II. DO INTERNET PLATFORMS' CONTENT MODERATION POLICIES RECOGNIZE NONPARTISANSHIP OR IMPARTIALITY AS A STATED PRINCIPLE?

Given the ongoing fervor to reform or repeal Section 230, one would expect that the internet platforms would undertake new steps to address the concerns raised by lawmakers, DOJ, and the Executive Order, particularly on the charge of political bias. However, Part II shows how the internet platforms' stated policies and practices do not adequately explain how they operationalize nonpartisanship as a principle of content moderation, either generally or specifically for candidates for public office.

A. Overview

All internet platforms engage in content moderation, meaning some amount of review of content posted by their users and removal of, or other remedial action for, content that violates the platform's community standards or guidelines.[420] Community standards are a euphemism for the rules that identify the kinds of content users are not permitted to share on the platform.[421] These community

418. *Id.*

419. Makena Kelly, *Joe Biden Wants to Revoke Section 230*, VERGE (Jan. 17, 2020, 10:29 AM), https://www.theverge.com/2020/1/17/21070403/joe-biden-president-election-section-230-communications-decency-act-revoke.

420. *See* GILLESPIE, *supra* note 1, at 5 ("There is no platform that does not impose rules, to some degree."); SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA 27–28 (2019) ("[T]he content is subject to an ecosystem made up of intermediary practices, policies, and people . . . .").

421. *See* GILLESPIE, *supra* note 1, at 45–46 (discussing the use of community standards as rules for users).

standards are de facto speech codes. If a user post violates a standard and the internet platform catches it, the violating content is removed or moderated. Content moderation lies in tension with the perception of an "open" platform.[422] When people disagree with content moderation, they often call it "censorship"—as evident in the titles of the Executive Order and three of the Section 230 reform bills.[423] Yet content moderation is essential. As Tarleton Gillespie writes: "Platforms must . . . moderate: both to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal—as well as to present their best face to new users, to their advertisers and partners, and to the public at large."[424] However, people's ambivalence to content moderation may leave internet platforms in a Catch-22: attacked for moderating too much and too little.

If we examine the community standards of internet platforms posted on their websites, they say little, if anything, about their treatment of politicians, let alone how, if at all, their content moderation of politicians and political campaign ads adheres to nonpartisanship. More generally, the community standards typically do not provide much specific detail about the precise procedures, mechanics, guiding principles, or timetable that content moderators must follow.[425] And none appears to disclose whether high-level executives, such as the CEO, can veto or override the decisions of content moderators who flagged a violation—and if so, by what criteria.[426] It is unclear whether internet platforms have any internal documents or handbooks setting forth such information for their moderators. Indeed, more details about the actual content moderation process arguably can be found in Kate Klonick's *Harvard Law Review* article than on the internet platforms' websites.[427]

---

422. *Id.* at 5.

423. *See supra* notes 352, 359, 364, 392 and accompanying text.

424. GILLESPIE, *supra* note 1, at 5.

425. *See, e.g.*, *Our Approach to Policy Development and Enforcement Philosophy*, TWITTER, https://help.twitter.com/en/rules-and-policies/enforcement-philosophy [https://perma.cc/R9XF-TJE2] (listing only five factors that Twitter says it considers when deciding when to take enforcement action).

426. *See, e.g.*, *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/introduction (discussing some of the policy choices that go into Facebook's Community Standards, but not mentioning executive veto power).

427. *See* Klonick, *supra* note 61, at 1639–42 (discussing Facebook's tiered content moderation system).

Below is a summary of the community standards of the large internet platforms. This summary shows that each platform has noticeable gaps in explaining how, if at all, it maintains nonpartisanship or impartiality in the content moderation of political candidates or political campaign ads. An important caveat is that the companies' policies may change, with refinements and updates. Indeed, during the writing of this Article, the policies of several companies changed, making it challenging to review a moving target. For example, both Facebook and Twitter announced that they were rolling back some of the new measures to stop misinformation they implemented during the 2020 election.[428] (For reference, the policies were last examined in December 2020.)

428. *See* Sarah Frier, *Facebook, Twitter Reverse Changes Meant to Curb Vote Misinformation*, BLOOMBERG (Dec. 16, 2020, 10:08 PM), https://www.bloomberg.com/news/articles/2020-12-17/facebook-twitter-undo-changes-meant-to-curb-vote-misinformation.

Table 1. Comparison of Content Moderation Policies

Company	Policy recognizes nonpartisanship or impartiality	Policy describes steps in review process	Penalties for violations	Public Interest Newsworthiness Exception	Right of Appeal to User
Twitter	Strives for "uniform consistency."	Some on public interest exception, but not much on violation decisions	1. tweet less visible 2. hiding tweet while awaiting removal 3. requiring user to remove tweet 4. stop direct messages 5. disabling account 6. read-only mode 7. permanent suspension	Yes	Yes
Facebook	Apply "consistently and fairly." IFCN "nonpartisanship, fairness" in fact-checking	Some	1. remove content 2. warning over content 3. disable account 4. escalation to external agencies	Yes	Yes
YouTube	Apply "consistently, without regard to a video's political viewpoint"	Some	1. remove content 2. age-restrict content 3. 3-strikes in 90 days results in termination	Yes	Yes
Reddit	No	Some in 2020 Security Report	1. request to stop 2. temporary or permanent suspension of accounts 3. removal of privileges from, or adding restrictions to, accounts 4. adding restrictions to communities, such as adding NSFW tags or Quarantining 5. Removal of content 6. Banning of Reddit communities	Yes	Yes
Snapchat	Apply "to all … equally"	No	remove the offending content, terminate your account, and/or notify law enforcement.	Yes	Unclear but seems not
Twitch	"[E]qual protections" v. hateful conduct	Some re: Hateful Conduct policy	1. removal of content 2. strike on the account 3. suspension, temporary and permanent	No	Yes
TikTok	Apply "to everyone … and everything"	Some in 2020 Trans. Report	1. removal of content 2. suspend or ban accounts 3. report to authorities	Yes	Yes

B. Twitter

Twitter has sparked the most controversy in moderating President Trump's tweets for violating its community standards.[429] A strength of Twitter's policy is its detailed and organized explanation. Twitter has numerous pages detailing its community standards.[430] Indeed, Twitter has the most detailed explanation of its community standards of all the companies surveyed. In a very helpful post, Twitter explains, at length, its "approach to policy development and enforcement philosophy."[431] Twitter professes that its rules are meant "to help ensure everyone feels safe expressing their beliefs and we strive to enforce them with uniform consistency."[432] Twitter says it "empower[s] people to understand different sides of an issue and encourage[s] dissenting opinions and viewpoints to be discussed openly."[433] Twitter devotes a page to describing its "civic integrity policy" to protect against "attempts to use our services to manipulate or disrupt civic processes," such as elections.[434] A separate page explains Twitter's policy on deceptively "synthetic and manipulated media."[435]

One unique feature of Twitter's remedial action: except for repeat offenders or "egregious" violations, Twitter does not remove violating tweets. Instead, Twitter asks "violators to remove the Tweet(s)" and may take "additional actions like verifying account ownership and/or temporarily limiting their ability to Tweet for a set period of time."[436] Egregious violations "result in the immediate and permanent suspension of an account."[437] Users can appeal enforcement decisions

---

429. *See supra* note 4 and accompanying text.

430. *See Twitter Rules and Policies*, TWITTER, https://help.twitter.com/en/rules-and-policies#twitter-rules [https://perma.cc/E33F-7K3P] (linking to Twitter's numerous community standard pages).

431. *Our Approach to Policy Development and Enforcement Philosophy*, *supra* note 425.

432. *Id.*

433. *Id.*

434. *Civic Integrity Policy*, TWITTER, https://help.twitter.com/en/rules-and-policies/election-integrity-policy [https://perma.cc/Y7VX-2BKL].

435. *Synthetic and Manipulated Media Policy*, TWITTER, https://help.twitter.com/en/rules-and-policies/manipulated-media [https://perma.cc/R9XC-7G45].

436. *Our Approach to Policy Development and Enforcement Philosophy*, *supra* note 425.

437. *Id.*; *see also Our Range of Enforcement Options*, TWITTER, https://help.twitter.com/en/rules-and-policies/enforcement-options [https://perma.cc/R3K3-KKQY] ("When we determine that a Tweet violated the Twitter Rules, we require the violator to remove it before they can Tweet again. We send an email notification to the violator identifying the Tweet(s) in violation and which policies have been

at Twitter.[438] Starting in October 2019, Twitter banned paid political ads, as announced in a tweet by CEO Jack Dorsey.[439] Twitter's transparency report of its content moderation from January to June of 2019 does not appear to contain specific categories for moderation of politicians, voter suppression, or election misinformation.[440] However, it does indicate "a 32% increase in the number of accounts actioned for violations of our civic integrity policy during this reporting period."[441] After the 2020 election, Twitter announced that it had added informational labels to 300,000 tweets about the presidential election, which represented a mere 0.2 percent of all tweets about the election.[442]

Despite the extensiveness of Twitter's policy, noticeably absent from Twitter's explanation of its "enforcement philosophy" is a discussion of the actual enforcement procedures—including any safeguards against bias or partisanship—that are used to determine whether a tweet violates its rules. This omission from its website is Twitter's biggest deficiency. The omission starkly contrasts with Twitter's lengthy discussion of its remedial actions once a violation has been found.[443] Twitter's lack of transparency on the procedures for determining violations of its Twitter rules has exposed it to criticism, such as with respect to its termination of 7,000 QAnon accounts for alleged violations.[444] In an April 1, 2020 post related to how Twitter was dealing with COVID-19, Twitter disclosed that it was relying more heavily on automated review of tweets, but that it will not impose permanent suspension of accounts without human review.[445] However, Twitter fails to outline the steps and composition

---

violated. They will then need to go through the process of removing the violating Tweet or appealing our review if they believe we made an error.").

438. *Our Range of Enforcement Options*, *supra* note 437.

439. jack (@jack), TWITTER (Oct. 30, 2019, 4:05 PM), https://twitter.com/jack/status/1189634360472829952.

440. *See* TWITTER, RULES ENFORCEMENT, https://transparency.twitter.com/en/reports/rules-enforcement.html#2019-jul-dec [https://perma.cc/UQ9P-KHB2].

441. *Id.*

442. *See* Kate Conger, *Twitter Says It Labeled 0.2% of All Election-Related Tweets as Disputed*, N.Y. TIMES (Nov. 12, 2020), https://www.nytimes.com/2020/11/12/technology/twitter-says-it-labeled-0-2-of-all-election-related-tweets-as-disputed.html.

443. *See Our Range of Enforcement Options*, *supra* note 437.

444. *See* Douek, *supra* note 335.

445. @Vijaya & Matt Derella, *An Update on Our Continuity Strategy During COVID-19*, TWITTER (Mar. 16, 2020), https://blog.twitter.com/en_us/topics/company/

of its review of content for violations of its Twitter rules—e.g., what percentage of tweets are moderated by automated process, when do human reviewers get involved, how many human reviewers typically review a tweet before it is determined to be a violation, and can high-level executives veto and override the decisions of content moderators?

To appreciate the lack of transparency in Twitter's process for determining a violation, one only needs to compare it with Twitter's quite detailed page explaining its "public-interest exception."[446] Under this exception, Twitter may take less severe remedial measures in response to a tweet than requiring its removal if the public interest so warrants.[447] This exception may apply to content of elected and government officials that violate the Twitter rules.[448] As shown in Table 1 above, all of the internet platforms, except Twitch, recognize a public interest exception of some kind.[449] Twitter outlines a four-step process for how it determines whether to apply a public-interest exception to an offending tweet of a political candidate.[450] I have inserted labels in brackets identifying the different persons conducting each step of the review process.

> 1. Our global enforcement team will escalate any Tweet that meets the criteria defined above for secondary review by our Trust & Safety team. We will not evaluate Tweets for the public interest exception if they do not violate the Twitter Rules or otherwise fail to meet the criteria above. [*Global Enforcement team*]
>
> 2. Our Trust & Safety team will evaluate the Tweet and prepare a recommendation on whether or not continued access to the Tweet is in the public interest. [*Trust & Safety team*]
>
> 3. The recommendation will be shared with a cross-functional set of leaders across different internal teams with diverse and multidisciplinary backgrounds in government, human rights, journalism, news, technology, and law, as well as in-market teams with an understanding of the cultural context in which the Tweet was posted. [*Cross-functional stakeholders*]

2020/An-update-on-our-continuity-strategy-during-COVID-19.html [https://perma.cc/X7AM-M2P3].

446. *About Public-Interest Exceptions on Twitter*, *supra* note 49.

447. *Id.*

448. *See id.* (listing an account's connection to a "current or potential member of a local, state, national, or supra-national governmental or legislative body" as a necessary requirement to apply Twitter's public interest exception).

449. *See supra* Table 1.

450. *About Public-Interest Exceptions on Twitter*, *supra* note 49.

> 4. After informing these cross-functional stakeholders of the recommendation and feedback from the cross-functional team, senior leaders from Trust & Safety will make the final decision to remove the Tweet or apply the notice.[451] [*Senior leaders of Trust & Safety team*]

Unlike Facebook, Twitter does "not consult externally on individual enforcement decisions."[452] Twitter also notes: "As with any enforcement action, our goal is *consistent* and transparent application of the notice, taking into account local context."[453] The page also includes helpful examples or situations in which it is likely to apply (or not) the public interest exception.[454]

Twitter relies on three internal groups to evaluate whether to invoke the public interest exception as a remedy to a violation of the community standard by a political candidate: (1) the frontline Global Enforcement team, (2) the Trust & Safety team, and (3) what Twitter characterizes as "cross-functional stakeholders," i.e., "leaders across different internal teams with diverse and multidisciplinary backgrounds."[455] The use of multilayers of review, in which a content moderation decision can be "escalate[d]" up the chain of review, is a typical approach of large internet platforms.[456] Twitter lists several options for penalties, ranging from removal or making less visible the violating tweet to temporary or permanent suspension of an account.[457] Twitter allows the user to appeal a violation determination.[458]

Although extensive, Twitter's explanation of its "public interest" policy has noticeable gaps. First, Twitter's discussion is framed around the determination of the "public interest" exception (to the remedy or moderation), but not the underlying violation. In other words, how does the global enforcement team determine if there is a violation in the first place, before the team escalates it? Detailed

---

451. *Id.*
452. *Id.*
453. *Id.* (emphasis added).
454. *Id.*
455. *Id.*
456. *See* Klonick, *supra* note 61, at 1647–48 (explaining how Facebook's Tier 3 content moderators can escalate review of content violating Facebook's community standards to Tier 2 moderators, and how Twitter uses "specialized team members" to review culturally-specific content previously flagged for removal).
457. *See Our Approach to Policy Development and Enforcement Philosophy*, *supra* note 425 (tying severity of punishment to the repetition and perceived egregiousness of the violation).
458. *See Our Range of Enforcement Options*, *supra* note 437 (linking to a platform interface to file appeals).

information about the process by which Twitter determines whether a violation has occurred would be equally as helpful. Moreover, it is unclear whether Twitter's "secondary review" by the Trust & Safety team or the input of the cross-functional stakeholders can recommend a reversal of the underlying violation determination. Assuming they can, it is also unclear how, if at all, nonpartisanship or impartiality is ensured during the internal process to determine violations. Twitter indicates that the *source* of the content is considered in deciding whether to apply the public interest exception, so the apparent lack of anonymity in the determination might create a risk of bias against the source.[459] Although the page does reiterate "our goal is consistent and transparent application of the [public interest] notice,"[460] consistency in the application of the public interest exception after a violation has already been found does not necessarily ensure that the determination of a violation was nonpartisan.

C. Facebook

Facebook's content moderation policy shares some of the same strengths and weaknesses as Twitter's. Starting in April 2018,[461] Facebook has published its internal enforcement guidelines, now referred to as its community standards.[462] Like Twitter, Facebook provides a landing page for its community standards that contains links to more detailed pages about the individual standards.[463] Facebook includes a page for "Understanding the Community Standards Enforcement Report" that explains not only how it prepares its transparency reports for content moderation but also

---

459. *See Our Approach to Policy Development and Enforcement Philosophy*, *supra* note 425. ("Some of the factors that help inform our decision-making about content are the impact it may have on the public, the source of the content, and the availability of alternative coverage of an event.").

460. *About Public-Interest Exceptions on Twitter*, *supra* note 49.

461. *See* Monika Bickert, *Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, FACEBOOK (Apr. 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards [https://perma.cc/Z5QC-N8FJ] (explaining that Facebook has upheld internal community standards "[f]or years").

462. *Community Standards*, *supra* note 426.

463. *See id.* (providing links to Facebook community standards for "[v]iolence and criminal behavio[]r," "[s]afety," "[o]bjectionable content," "[i]ntegrity and authenticity," "[r]especting intellectual property," and "[c]ontent-related requests").

how it detects possible violations and reviews them.[464] Facebook provides more description than Twitter on this important issue. Facebook states:

> We use technology, human review or a combination of the two to determine whether a piece of content violates our policies. If the content is routed to our human review team, then they use *our policies and a step-by-step process to help them make decisions accurately and consistently for the appropriate violation type*. We also provide our reviewers with tools to review the reported content and the available context required to identify the concern and determine whether a piece of content violates a standard.[465]

However, Facebook does not provide specific details about the "step-by-step process" it refers to.[466] Disclosing this step-by-step process, ideally with a diagram of each step, would be helpful to understand how Facebook determines violations. For example, how precisely does Facebook's step-by-step process promote consistent or accurate decisions of violations? Moreover, what "tools to review the reported content" does Facebook give to its reviewers?

Like Twitter, Facebook uses a range of penalties, from a warning to removal of content, and recognizes a right of appeal for a violation decision.[467] In the introduction to its community standards, Facebook recognizes a public interest or newsworthiness exception, similar to Twitter's, by which Facebook "allow[s] content that would otherwise go against our Community Standards—if it is newsworthy and in the public interest."[468]

Because Facebook's explanations of its policy are scattered on various pages, some even outside of its community standards, the average user is likely to have a harder time finding Facebook's policy compared to Twitter's well-organized layout. For example, Facebook's policies regarding election integrity and political ads are not located in or findable on the community standards landing page. Instead, most of Facebook's policies on these important areas are scattered in its

---

464. *Understanding the Community Standards Enforcement Report*, FACEBOOK, https://transparency.facebook.com/community-standards-enforcement/guide.

465. *Id.* (emphasis added).

466. *See id.* (explaining only that, in the event content is routed to human review, such content is reviewed using Facebook's policies and a "step-by-step process").

467. *See id.* (listing as possible actions "removing content," "covering content with a warning," "disabling accounts," and "escalations to external agencies").

468. *Community Standards*, *supra* note 462.

newsroom blog,[469] Business Help Center,[470] and even Zuckerberg's personal Facebook profile.[471] This scattered placement makes it very difficult to determine precisely Facebook's policy on the important issues of election integrity, misinformation, and paid political ads.

Another complicating factor is that Facebook's policy changed quite dramatically in 2020, but there is no easy way for users to track all the changes. Twitter has the same problem. Ideally, the platforms would provide a fixed page that lists all changes to their content moderation policies by date, as YouTube does for its COVID-19 misinformation policy.[472]

On the controversial issue of whether Facebook will fact-check politicians and review their content for violations of its community standards, it is difficult to identify a clear statement of the policy. At some point during the 2020 election, Facebook started fact-checking the content of politicians for election misinformation. This marked a dramatic change for Facebook. Facebook's prior approach was hands-off. In June 2020, Facebook's Business Help Center page explained the exception Facebook gave at the time to politicians, exempting their posts and political ads from any fact-checking.[473] But that page

---

469. *See, e.g.*, Katie Harbath & Samidh Chakrabarti, *Expanding Our Efforts to Protect Elections in 2019*, FACEBOOK (Jan. 28, 2019), https://about.fb.com/news/2019/01/elections-2019 [https://perma.cc/UJR9-7BV6] (explaining Facebook's "multifaceted" approach to protecting election integrity by blocking or removing fake accounts, identifying and removing "bad actors," limiting fake news, and providing transparency); Guy Rosen et al., *Helping to Protect the 2020 US Elections*, FACEBOOK (Oct. 21, 2019), https://about.fb.com/news/2019/10/update-on-election-integrity-efforts [https://perma.cc/W8RR-GTSU] (highlighting, among others, Facebook's efforts to fight foreign election interference).

470. *About Ads About Social Issues, Elections or Politics*, FACEBOOK (July 2, 2020), https://www.facebook.com/business/help/167836590566506 [https://perma.cc/9SFH-3Z4X]; *Prohibited Ads About Social Issues, Elections or Politics in the United States and Information on the 2020 Restriction Period*, FACEBOOK (Oct. 7, 2020), https://www.facebook.com/business/help/253606115684173 [https://perma.cc/2F82-YEHC]; *Fact-Checking on Facebook*, FACEBOOK (Aug. 12, 2020), https://www.facebook.com/business/help/182222309230722 [https://perma.cc/BV4V-DPYL].

471. *See* Zuckerberg, *supra* note 9 (highlighting Facebook's new policies towards informing users about voting, preventing voter suppression, and combatting hate speech).

472. *See Coronavirus Disease 2019 (COVID-19) Updates*, YOUTUBE, https://support.google.com/youtube/answer/9777243?hl=en&ref_topic=6151248 [https://perma.cc/6YEP-YCD3].

473. *See Over 130 Companies Remove Ads from Facebook in #StopHateforProfit Boycott, Forcing Mark Zuckerberg to Change Lax Facebook Policy on Misinformation and Hate Content*, FREE INTERNET PROJECT (June 27, 2020) [hereinafter *130 Companies Remove Ads*], https://thefreeinternetproject.org/blog/over-130-companies-remove-ads-facebook-

has since been removed and the Facebook "Fact-Checking" page now omits any mention of treatment of political ads or politicians.[474] Earlier in 2020, Zuckerberg defended Facebook's prior approach: "I just believe strongly that Facebook shouldn't be the arbiter of truth of everything that people say online."[475] In September 2019, Facebook Vice President Nick Clegg had explained this prior policy in a blog post: "[F]rom now on we will treat speech from politicians as newsworthy content that should, as a general rule, be seen and heard."[476] However, politicians' ads and posts were still subject to other aspects of Facebook's community standards, including hate speech and incitement of violence.[477] In June 2020, Facebook removed ads for the Trump campaign that included a symbol (red triangle) associated with a Nazi symbol, for example.[478]

Then, on June 26, 2020, Zuckerberg announced substantial changes to Facebook's policy after over 400 companies boycotted the company by removing their ads from Facebook to protest the "hate, bias, and discrimination growing on [Facebook's] platforms."[479] On his personal Facebook page, Zuckerberg said that Facebook would add labels to any violation of Facebook's community standards that remain on Facebook under the newsworthiness exception, including violations by political candidates.[480] Zuckerberg clarified that there is no newsworthiness exception for "[c]ontent that incites violence or

---

stophateforprofit-boycott-forcing-mark-zuckerberg-change [https://perma.cc/DHZ2-4ZMQ] (describing Facebook's since-reversed policy not to fact-check political ads).

474. *See Fact-Checking on Facebook*, *supra* note 470.

475. Rebecca Klar, *Zuckerberg: 'Facebook Shouldn't Be the Arbiter of Truth of Everything that People Say Online,'* HILL (May 27, 2020, 8:46 PM), https://thehill.com/policy/technology/499852-zuckerberg-facebook-shouldnt-be-the-arbiter-of-truth-of-everything-that [https://perma.cc/9FHB-M5AM].

476. Clegg, *supra* note 100.

477. *See* Zuckerberg, *supra* note 9 (asserting Facebook's commitment to remove content that is hate speech, incites violence, or suppresses voting, regardless of the source).

478. Donie O'Sullivan, *Facebook Says It Took down Trump Ads Because They Used Nazi Symbol*, CNN (June 19, 2020, 5:42 AM), https://www.cnn.com/2020/06/18/tech/facebook-trump-ads-triangle-takedown/index.html [https://perma.cc/L8RR-VZJR].

479. *Calling on Facebook Corporate Advertisers to Pause Ads for July 2020*, COLOR OF CHANGE (June 19, 2020), https://colorofchange.org/stop-hate-for-profit [https://perma.cc/VPX3-8XLN].

480. *See* Zuckerberg, *supra* note 9 ("We will soon start labeling some of the content we leave up because it is deemed newsworthy . . . .").

suppresses voting," including in posts by political candidates.[481] Facebook tightened its enforcement of voter suppression through false claims about polling conditions, as well as its "ads policy to prohibit claims that 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."[482] Zuckerberg commented, "There are no exceptions for politicians in any of the policies that I'm announcing here today."[483]

Zuckerberg announced even greater changes on his Facebook page on September 3, 2020, which included nearly a dozen major changes or initiatives Facebook was taking to combat election misinformation, including related to election results.[484] These measures do not appear to be incorporated into Facebook's community standards. Instead, Zuckerberg's post was embedded into a Facebook news release, which is also included in the newsfeed for Facebook's special page "Preparing for Elections."[485] In any event, three of the changes are worth noting. First, Facebook banned political ads starting the week before the election.[486] Second, Facebook said it would remove implicit

---

481. Catherine Thorbecke, *Facebook to Label 'Newsworthy' Posts that Violate Rules as Ad Boycotts Grow*, ABC NEWS (June 26, 2020, 4:48 PM), https://abcnews.go.com/Business/facebook-label-newsworthy-posts-violate-rules-ad-boycotts/story?id=71478554 [https://perma.cc/TE5X-QJUQ].

482. Zuckerberg, *supra* note 9.

483. Shannon Bond, *In Reversal, Facebook to Label Politicians' Harmful Posts as Ad Boycott Grows*, NPR (June 26, 2020, 2:05 PM), https://www.npr.org/2020/06/26/883941796/unilever-maker-of-dove-soap-is-latest-brand-to-boycott-facebook [https://perma.cc/CRP4-EXZV].

484. *See New Steps to Protect the US Elections*, FACEBOOK (Sept. 3, 2020), https://about.fb.com/news/2020/09/additional-steps-to-protect-the-us-elections [https://perma.cc/7UV8-5WTN] (announcing Facebook's policy to add labels to any candidate's premature declaration of electoral victory on the platform); *Mark Zuckerberg: Facebook to Suspend Political Ads Week Before US Election, Add a Label to Premature Election Claims of Victory*, FREE INTERNET PROJECT (Sept. 3, 2020), https://thefreeinternetproject.org/blog/mark-zuckerberg-facebook-suspend-political-ads-week-us-election-add-label-premature-election [https://perma.cc/S2HU-SR6G] (describing each of Zuckerberg's September 3, 2020 Facebook policy changes).

485. *Preparing for Elections*, FACEBOOK, https://about.fb.com/actions/preparing-for-elections-on-facebook [https://perma.cc/2LGR-GT6Z].

486. *5 Things to Remember About Political and Issue Advertising Around the US 2020 Election*, FACEBOOK (Oct. 26, 2020), https://www.facebook.com/business/news/facebook-ads-restriction-2020-us-election [https://perma.cc/52H3-JRRT]. Facebook allowed political ads for the runoff election for the Senate races in Georgia. *See* Emily Glazer & Patience Haggin, *Facebook to Allow Political Ads for Georgia Runoffs*, WALL ST. J.

misrepresentations about the voting process just as it has removed explicit misrepresentations.[487] Third, Facebook announced a new policy of adding "informational label[s]" to posts on Facebook that attempt to "delegitimize the outcome of the election" such as "claiming that lawful methods of voting will lead to fraud."[488] As it turned out, President Trump himself was a main source of false claims of election results that warranted Facebook's informational labels, although one study suggested that the labels had only a modest effect on reducing the sharing of the flagged posts on Facebook.[489]

Facebook has an Elections Operation Center to combat election misinformation, but there is no dedicated webpage for it.[490] Facebook does include a page for its Facebook Protect program for the security of verified accounts of political candidates.[491] But this special program for political candidates is untethered from the content moderation policies that political candidates are expected to follow. Facebook issues a transparency report of its content moderation, but its report in November 2020 does not appear to identify content moderation involving political ads, content of political candidates, election-related misinformation, or voter suppression.[492] However, after the 2020 election, Facebook stated that it had added informational labels to more than 180 million posts containing election misinformation between March and November 3, 2020.[493]

(Dec. 15, 2020, 4:38 PM), https://www.wsj.com/articles/facebook-to-allow-political-ads-for-georgia-runoffs-11608062846.

487. Mike Isaac, *Facebook Moves to Limit Election Chaos in November*, N.Y. TIMES (Sept. 22, 2020), https://www.nytimes.com/2020/09/03/technology/facebook-election-chaos-november.html.

488. *New Steps to Protect the US Elections*, *supra* note 484.

489. *See* Craig Silverman & Ryan Mac, *Facebook Knows that Adding Labels to Trump's False Claims Does Little to Stop Their Spread*, BUZZFEED NEWS (Nov. 16, 2020, 8:07 PM), https://www.buzzfeednews.com/article/craigsilverman/facebook-labels-trump-lies-do-not-stop-spread [https://perma.cc/B6YR-FFNP].

490. *See* Rosen et al., *supra* note 469 (describing abstractly the Elections Operation Center's efforts to remove content that interferes with or suppresses voting).

491. *Facebook Protect*, FACEBOOK, https://www.facebook.com/gpa/facebook-protect [https://perma.cc/UUY4-NR9V].

492. *See* FACEBOOK, COMMUNITY STANDARDS ENFORCEMENT REPORT, https://transparency.facebook.com/community-standards-enforcement (last visited Jan. 28, 2021) (omitting these categories from covered policy areas in November 2020 report).

493. *See* Danielle Abril, *Facebook Reveals that Massive Amounts of Misinformation Flooded Its Service During the Election*, FORTUNE (Nov. 19, 2020, 2:49 PM), https://fortune.com/2020/11/19/facebook-misinformation-labeled-180-million-posts-2020-election-hate-speech-prevalence [https://perma.cc/34B6-9PYL].

Facebook also removed 265,000 posts on Facebook and Instagram as containing voter suppression efforts.[494]

Finally, Facebook says it "work[s] to apply these policies in a way that is fair and consistent to all communities and cultures around the world,"[495] and in a way that is "inclusive of different views and beliefs."[496] Since 2016, Facebook has "partner[ed] with independent third-party fact-checkers globally who are certified through the non-partisan International Fact-Checking Network (IFCN)."[497] IFCN has a Code of Principles that members are to abide by; the first principle is "[a] commitment to Non-partisanship and Fairness," with five criteria for members who fact-check to follow.[498] The IFCN Code of Principles is a step in the right direction. But the Code's status at Facebook is ambiguous. They are not mentioned in Facebook's community standards. Instead, they are located on a separate "Journalism Project," in which Facebook describes what its partner fact-checkers are supposed to abide by.[499] It is unclear whether they also apply to decisions of Facebook employees. Assuming they do, it is still unclear the extent to which the Code of Principles applies to fact-checking the content of politicians.[500] Facebook did *not* fact-check "[p]osts and ads from politicians" under Facebook's policy through the summer of 2020, which apparently was modified by Zuckerberg's September 3, 2020 post to extend some fact-checking to politicians' content.[501] In any event, IFCN itself has faced criticism for having

494. *Id.*

495. *Understanding the Community Standards Enforcement Report*, *supra* note 464.

496. *Community Standards*, *supra* note 426.

497. *Facebook's Approach to Misinformation: Partnering with Third-Party Fact-Checkers*, FACEBOOK JOURNALISM PROJECT, https://www.facebook.com/journalismproject/programs/third-party-fact-checking/selecting-partners [https://perma.cc/HY5L-N6NL].

498. *The Commitments of the Code of Principles*, IFCN, https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles [https://perma.cc/BLF7-JFR9].

499. *See Facebook's Approach to Misinformation: Partnering with Third-Party Fact-Checkers*, *supra* note 497 (requiring partner fact-checkers to commit to "[n]onpartisanship and [f]airness"; "[t]ransparency of [s]ources[,] . . . [f]unding[,] . . . [and m]ethodology"; and "[o]pen and [h]onest [c]orrections").

500. *See About*, IFCN, https://ifcncodeofprinciples.poynter.org/know-more [https://perma.cc/7LJ2-DQU4] (delimiting the application of the principles to organizations that "regularly publish nonpartisan reports on the accuracy of statements").

501. *Compare New Steps to Protect the US Elections*, *supra* note 484 (describing Facebook's post-September 3, 2020 content policy), *with 130 Companies Remove Ads*, *supra* note 473 (describing Facebook's pre-September 3, 2020 content policy).

members whose fact-checking is politically biased.[502] Moreover, Facebook appears to allow high-level executives to intervene to override or reverse the violation determinations by content moderators.[503] According to NBC News, "Facebook employees in the misinformation escalations team, with direct oversight from company leadership, deleted strikes during the review process that were issued to some conservative partners for posting misinformation over the last six months."[504] A similar controversy arose in India where, according to the *Wall Street Journal*, Facebook's "top public-policy executive in the country, Ankhi Das, opposed applying the hate-speech rules to Mr. Singh and at least three other Hindu nationalist individuals and groups flagged internally for promoting or participating in violence."[505] (Facebook denied the allegation, without discussing Das's involvement.[506])

In short, other than a nod to independent IFCN fact-checkers, Facebook fails to explain how nonpartisanship is operationalized or how political bias is avoided during its content moderation of political candidates or political ads, or the creation of its moderation policy (which reportedly was designed with a special concern to avoid negatively affecting conservative content).[507] Such "step-by-step" mechanisms may exist at Facebook, but they are not disclosed.

---

502. *See* Anton Troianovski, *Fighting False News in Ukraine, Facebook Fact Checkers Tread a Blurry Line*, N.Y. TIMES (July 26, 2020), https://www.nytimes.com/2020/07/26/world/europe/ukraine-facebook-fake-news.html ("Stopfake[, one of Facebook's outside fact-checkers,] has been battling accusations of ties to the Ukrainian far right and of bias in its fact-checking."); Nandini Jammi, *How Did the Daily Caller Become a Facebook Fact-Checker?*, MEDIUM (Oct. 30, 2019), https://medium.com/@nandoodles/how-did-the-daily-caller-become-a-facebook-fact-checker-2a2dd7042c4f [https://perma.cc/77XT-DWV2] (highlighting The Daily Caller's "ties to white supremacists").

503. *See* Solon, *supra* note 34; Newley Purnell & Jeff Horwitz, *Facebook's Hate-Speech Rules Collide with Indian Politics*, WALL ST. J. (Aug. 14, 2020, 12:47 PM), https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346.

504. Solon, *supra* note 34.

505. Purnell & Horwitz, *supra* note 503.

506. *See* Sunil Prabhu, *Decisions not Unilateral: Facebook Defends India Policy Chief Ankhi Das*, NDTV (Sept. 3, 2020, 6:22 PM), https://www.ndtv.com/india-news/not-unilateral-facebook-responds-to-congress-on-policy-decisions-teams-2289950 [https://perma.cc/S2AN-PXPH] (reporting Facebook's position that "[e]nforcing policies on hate speech is 'not [decided] unilaterally by any one person'").

507. *See Solon*, *supra* note 34 (detailing how conservative Facebook pages "were not penalized for violations of the company's misinformation policies").

D. YouTube and Google

YouTube's community standards have some of the same weaknesses as Facebook's. Some important information related to content moderation of political ads and candidates is not contained in the community standards, but is instead scattered across YouTube's official blog and even pages on Google—making it hard to find.

YouTube's community standards have a landing page that organizes the standards by categories over several pages, with very helpful examples of potential violations.[508] Some information related to voter suppression, false claims on political candidate eligibility for office, and manipulated media is buried in the page for "Spam, Deceptive Practices & Scams Policies,"[509] which is perhaps not the most obvious category. Unlike Twitter and Facebook, YouTube did not implement a policy to address election misinformation until after the election on December 9, 2020, to address false content challenging the presidential election results.[510] YouTube "terminated over 8000 channels and thousands of harmful and misleading elections-related videos for violating our existing policies."[511] The community standards added this change: "Presidential Election Integrity: Content that advances false claims that widespread fraud, errors, or glitches changed the outcome of any past U.S. presidential election."[512] YouTube explained that it "remove[s] content that misleads people by alleging that widespread fraud or errors changed the outcome of the 2020 U.S. presidential election uploaded on or after December 9."[513] The move came after YouTube faced criticism

508. *See Community Guidelines*, YOUTUBE, https://www.youtube.com/howyoutubeworks/policies/community-guidelines [https://perma.cc/9L8P-WNRZ] (providing links to YouTube's community guidelines on videos involving "[s]pam and deceptive practices," "[s]ensitive content," "[v]iolent or dangerous content," and "[r]egulated goods").

509. *Spam, Deceptive Practices & Scams Policies*, YOUTUBE, https://support.google.com/youtube/answer/2801973?hl=en&ref_topic=9282365 [https://perma.cc/D8GL-LYWS].

510. *See Supporting the 2020 U.S. Election*, YOUTUBE (Dec. 9, 2020), https://blog.youtube/news-and-events/supporting-the-2020-us-election [https://perma.cc/2YHP-7MWP].

511. *Id.*

512. *Spam, Deceptive Practices & Scams Policies*, *supra* note 509.

513. *Id.*

for allowing Trump and the One American News Network to share videos making false claims of victory for Trump.[514]

Before this belated change, YouTube did not appear to have a specific community standard for election misinformation consisting of false claims of victory or false claims of voter fraud; instead, YouTube had categories for voter suppression, manipulated media, and false claims on candidate eligibility. YouTube's blog contains greater explanation of "How YouTube Supports Elections," and its removal policy for "election-related content that violates our policies" with five examples of violations.[515] YouTube's election blog post recognizes a principle of consistency in the enforcement of its rules: "As always, we enforce our policies consistently, without regard to a video's political viewpoint."[516] YouTube's community standards recognize the principle of "consistent" application, but they stop short of using the language in the blog post, "without regard to a video's political viewpoint."[517]

YouTube briefly explains how it detects potential violations through a combination of human and machine review,[518] "task[ing] over 10,000 people with detecting, reviewing, and removing content that violates our guidelines" and allowing users to flag content as well.[519] Without providing much detail of the process, YouTube describes how human reviewers decide violations and "strikes"

---

514. *See* Daisuke Wakabayashi, *Election Misinformation Continues Staying up on YouTube.*, N.Y. TIMES (Nov. 10, 2020), https://www.nytimes.com/2020/11/10/technology/election-misinformation-continues-staying-up-on-youtube.html; David Ingram, *YouTube Says It Wants 'Discussion' of Election Results, even when It's Been Debunked*, NBC NEWS (Nov. 13, 2020, 3:41 PM), https://www.nbcnews.com/tech/social-media/youtube-says-it-wants-discussion-election-results-even-when-it-n1247764 [https://perma.cc/4J9Q-HQMZ].

515. Leslie Miller, *How YouTube Supports Elections*, YOUTUBE (Feb. 3, 2020), https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html [https://perma.cc/2SF4-AAQE].

516. *Id.*

517. *How Do We Develop New Policies and Update Existing Ones?*, YOUTUBE, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#developing-policies [https://perma.cc/D8QQ-53F2].

518. *See How Does YouTube Identify Content that Violates Community Guidelines?*, YOUTUBE, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#detecting-violations [https://perma.cc/3GZD-Q3SN] (explaining how the combined approach allows YouTube to "detect problematic content at scale").

519. *Is There a Way for the Broader Community to Flag Harmful Content?*, YOUTUBE, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#flagging-content [https://perma.cc/CST7-JYBE].

committed by users.[520] Three strikes by a user in ninety days results in termination of the account.[521] The main penalties are removal of a video or age-restricting adult content.[522] YouTube allows the user to appeal a violation, and YouTube's "teams will re-review the decision."[523] YouTube also issues a transparency report of its content moderation, but the report does not appear to identify political ads, content of political candidates, or election-related misinformation.[524]

Similar to Twitter and Facebook, YouTube has a public interest or newsworthiness exception to its community standards. YouTube states: "We might allow videos that depict dangerous acts [if] they're meant to be educational, documentary, scientific, or artistic (EDSA)."[525] At a business conference in September 2019, CEO Susan Wojcicki told the audience that YouTube applies the EDSA exception to politicians.[526] Wojcicki explained that YouTube has an exception for "educational, documentary, scientific or artistic (EDSA)" videos that may remain on YouTube, despite violating a community standard.[527] A YouTube spokesperson later clarified that politicians are still subject to the same community standards, presumably as to what constitutes a violation.[528] Unfortunately, this policy with respect

---

520. *What Action Does YouTube Take for Content that Violates Community Guidelines?*, YOUTUBE https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#enforcing-policies [https://perma.cc/B9D5-B8CD].

521. *Id.*

522. *See id.*

523. *Id.*

524. *See* GOOGLE, YOUTUBE COMMUNITY GUIDELINES ENFORCEMENT, https://transparencyreport.google.com/youtube-policy/removals?hl=en [https://perma.cc/2QSF-FR4B] (listing child safety; spam; nudity of sexual content; violent or graphic material; promotion of violence and violent extremism; harmful or dangerous material; harassment and cyberbullying as grounds for video removal).

525. *Harmful or Dangerous Content Policy*, YOUTUBE, https://support.google.com/youtube/answer/2801964?hl=en [https://perma.cc/3DDP-CJJN].

526. *See* Nina Golgowski, *YouTube CEO Says Politicians Are Exempt from Content Rules*, HUFFPOST (Sept. 25, 2019, 6:54 PM), https://www.huffpost.com/entry/politicians-exempt-from-youtube-rules-ceo-says_n_5d8ba3c4e4b01c02ca627f81 [https://perma.cc/CHE3-R2Y9] ("When you have a political officer that is making information that is really important for the constituents to see, or for other global leaders to see, that is content that we would leave up because we think it's important for other people to see . . . .").

527. *Id.*

528. Steven Overly, *YouTube CEO: Politicians Can Break Our Content Rules*, POLITICO (Sept. 25, 2019, 6:40 PM), https://www.politico.com/story/2019/09/25/youtube-ceo-politicians-break-content-rules-1510919 [https://perma.cc/DN9X-N4HB].

to politicians cannot be found in YouTube's community standards. YouTube allows political ads and reviews them for clear violations of its rules against "'deep fakes' (doctored and manipulated media), misleading claims about the census process, and ads or destinations making demonstrably false claims that could significantly undermine participation or trust in an electoral or democratic process."[529]

Google's search engine is different from the other platforms discussed above in that Google's search engine is not social media and does not disseminate user-generated content. Therefore, Google does not moderate user-generated content like the other platforms.[530] As a result, Google's community standards and code of conduct online are focused on its employees.[531] One of the community guidelines asks Google employees not to "disrupt[] the workday to have a raging debate over politics."[532] Google has a dedicated page to elections, including "Protecting Elections Information Online."[533] Google also has a page outlining its election ads policy.[534] Google has a general ad policy against misrepresentation and some forms of misleading content.[535] Google explains how it verifies U.S. election

529. Scott Spencer, *An Update on Our Political Ads Policy*, GOOGLE (Nov. 20, 2019), https://www.blog.google/technology/ads/update-our-political-ads-policy [https://perma.cc/6CJH-98UA].

530. *See* Matt Southern, *Google Doesn't Treat User Generated Content Different from Main Content*, SEARCH ENGINE J. (May 19, 2020), https://www.searchenginejournal.com/google-doesnt-treat-user-generated-content-different-from-main-content/369168/#close [https://perma.cc/563X-SYJU] ("Google doesn't differentiate between content you wrote and content your users wrote. If you publish it on your site, we'll see it as content that you want to have published. And that's what we'll use for ranking [search results].").

531. *See Community Guidelines*, GOOGLE, https://about.google/community-guidelines [https://perma.cc/JMA7-PUJY] ("The following guidelines . . . apply when you're communicating in the workplace."); *Google Code of Conduct*, ALPHABET, https://abc.xyz/investor/other/google-code-of-conduct [https://perma.cc/Z3SR-T5Y7] ("We expect all of our employees and Board members to know and follow the Code [of Conduct].").

532. Community Guidelines, *supra* note 531.

533. *Protecting Elections Information Online*, GOOGLE, https://elections.google/#protecting-elections [https://perma.cc/SN6P-5G2M].

534. *See Political Content*, GOOGLE, https://support.google.com/adspolicy/answer/6014595?hl=en [https://perma.cc/M7ML-AGLF] (explaining Google's election ads policy extends to ads "for political organizations, political parties, political issue advocacy or fundraising, and individual candidates and politicians").

535. *See Misrepresentation*, GOOGLE, https://support.google.com/adspolicy/answer/6020955?hl=en&ref_topic=1626336 [https://perma.cc/N4JZ-WEBH] (noting that Google's misrepresentation policy covers "ads or destinations that deceive users by

advertising.[536] However, Google does not specifically address the extent to which it fact-checks political ads.[537] The *Wall Street Journal* reported that Google rejected ads related to Senator Graham and President Trump that made unsubstantiated claims.[538] It is unclear whether Google applies a "newsworthiness" exception to its elections ad review. Republican lawmakers have accused Google's search of being biased against conservative viewpoints, and a *Wall Street Journal* study revealed a murky process by which Google makes human adjustments to its search algorithms to alter its search.[539] Google is secretive about its search algorithms (which are protected trade secrets) and its process of adjusting search results.[540] Under questioning about anti-conservative bias, CEO Sundar Pichai testified in December 2018, "I can commit to you and I can assure you, we do it without regards to political ideology. Our algorithms do it with no notion of political sentiment."[541] Yet it is unclear how nonpartisanship is ensured in Google search, especially when human refinements are made by Google employees.

---

excluding relevant product information or providing misleading information about products, services, or businesses").

536. *See About Verification for Election Advertising in the United States*, GOOGLE, https://support.google.com/adspolicy/answer/9002729?hl=en&ref_topic=1316596 [https://perma.cc/C4G9-9RVG] (listing Google's verification requirements for running election ads on its platform).

537. *See id.* (indicating only that Google will use verification information merely to "verify your identity and eligibility to run election ads").

538. Patience Haggin & Emily Glazer, *Facebook, Twitter and Google Write Their Own Rules for Political Ads—and What You See*, WALL ST. J. (June 4, 2020, 11:00 AM), https://www.wsj.com/graphics/how-google-facebook-and-twitter-patrol-political-ads.

539. *See* Kirsten Grind et al., *How Google Interferes with Its Search Algorithms and Changes Your Results*, WALL ST. J. (Nov. 15, 2019, 8:15 AM), https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753 (noting Google seldom discloses when or why changes to its search algorithm are made and that Google has interfered with search results to a far greater degree than publicly acknowledged).

540. *Id.*

541. Alina Selyukh, *Google CEO Says He Leads 'Without Political Bias' in Congressional Testimony*, NPR (Dec. 11, 2018, 3:19 PM), https://www.npr.org/2018/12/11/675543073/google-ceo-says-he-leads-without-political-bias-in-congressional-testimony [https://perma.cc/TPJ6-E7BL].

E. Reddit

Reddit is a platform that allows wide-ranging discussion forums or communities called "subreddits."[542] Although Reddit has a reputation for no-holds-barred discussions,[543] the platform recognizes eight rules or community standards.[544] Individuals who create a subreddit, called "moderators" or "mods," exercise considerable discretion over removing, approving, and labeling content, according to the rules created for the subreddit, but they are supposed to establish "clear, concise, and consistent guidelines" for the group.[545] The moderators are tasked with enforcing the eight rules of Reddit.[546] This is a unique feature of Reddit, placing more direct responsibility of content moderation and enforcement of community standards on its users. Reddit has employee "admins," who have greater authority in content moderation or rule enforcement.[547] Some moderators have complained that the review process by admins is murky.[548] Reddit allows appeals of suspensions called the "normal appeal flow," but it is unclear whether an appeal is allowed for a removal or quarantining of content.[549]

In its 2020 Security Report, Reddit revealed that it uses automated review to detect content manipulation and scaled attacks by bots on

---

542. *See, e.g.*, *Subreddits*, REDDIT, https://www.reddit.com/subreddits [https://perma.cc/3U24-X3QZ].

543. For a history of Reddit's increasing content moderation, see u/spez, *Upcoming Changes to Our Content Policy, Our Board, and Where We're Going from Here*, REDDIT (June 5, 2020, 2:04 PM), https://www.reddit.com/r/announcements/comments/gxas21/upcoming_changes_to_our_content_policy_our_board [https://perma.cc/C7EN-X66R].

544. *See Reddit Content Policy*, REDDIT, https://www.redditinc.com/policies/content-policy [https://perma.cc/3D8N-QRSA].

545. *See Moderator Guidelines for Healthy Communities*, REDDIT, https://www.redditinc.com/policies/moderator-guidelines [https://perma.cc/5LBB-TH3T].

546. *See id.*

547. *See* Kim Renfro, *For Whom the Troll Trolls: A Day in the Life of a Reddit Moderator*, BUS. INSIDER (Jan. 13, 2016, 12:27 PM), https://www.businessinsider.com/what-is-a-reddit-moderator-2016-1 [https://perma.cc/K6KK-ZYS3].

548. *See* u/ggAlex, *The Mod Conversations that Went into Today's Policy Launch*, REDDIT (June 29, 2020, 11:58 AM), https://www.reddit.com/r/modnews/comments/hi3nkr/the_mod_conversations_that_went_into_todays [https://perma.cc/MBE2-QLUR].

549. *See* u/worstnerd, *Improved Ban Evasion Detection and Mitigation*, REDDIT (May 28, 2020, 5:34 PM), https://www.reddit.com/r/redditsecurity/comments/gsgg6k/improved_ban_evasion_detection_and_mitigation [https://perma.cc/E6UL-78ST].

Reddit, but human review for each report of abuse.[550] Reddit stated that its admins had removed over seventy-six million posts for content manipulation, including by malicious bots, which could have been intended to interfere with the U.S. election.[551]

Notably, a Reddit moderator can create a subreddit that is politically partisan and that excludes comments that do not conform to the political views or political party adopted in the subreddit's rules.[552] Although Reddit's explanation states that it allows "views across the political spectrum," the community standards do not discuss how Reddit handles content moderation of political candidates.[553] Reddit's ad policy limits political ads to federal office campaigns and subjects each ad to manual review; the general ad policy against "deceptive, untrue, or misleading advertising" also applies to political ads.[554] Reddit recognizes "contextual exceptions" for live video that "technically break these policies, but are nonetheless important,"[555] and also appears to have a general newsworthiness exception for content beyond live videos.[556]

Due to criticisms for allowing hate speech directed at persons of color amidst the nationwide protests of the police killing of George Floyd,[557] Reddit beefed up its policy against hate speech in June 2020

---

550. *See* u/worstnerd, *Reddit Security Report—June 18, 2020*, REDDIT (June 18, 2020, 12:15 PM), https://www.reddit.com/r/redditsecurity/comments/hbiuas/reddit_security_report_june_18_2020 [https://perma.cc/DK3Y-5EWC].

551. *See id.*

552. *See, e.g.*, *r/Republican*, REDDIT, https://www.reddit.com/r/Republican [https://perma.cc/2C2B-YHBL]; *r/democrats/rules*, REDDIT, https://www.reddit.com/r/democrats/about/rules (asking users not to "promote other political parties" or "post material that is anti-Democrat").

553. *See Update to Our Content Policy*, *supra* note 7; *Reddit Content Policy*, *supra* note 544.

554. *Changes to Reddit's Political Ads Policy*, REDDIT (Apr. 13, 2020, 4:35 PM), https://www.reddit.com/r/announcements/comments/g0s6tn/changes_to_reddits_political_ads_policy [https://perma.cc/5CNF-32ET].

555. *Reddit Content Policy for Live Video*, REDDIT, https://www.redditinc.com/policies/broadcasting-content-policy [https://perma.cc/5QJK-YTDR].

556. *See, e.g.*, *Do Not Post Violent Content*, REDDIT, https://www.reddithelp.com/hc/en-us/articles/360043513151 [https://perma.cc/GD9D-JSSZ] ("We understand there are sometimes reasons to post violent content (e.g., educational, newsworthy, artistic, satire, documentary, etc.) so if you're going to post something violent in nature that does not violate these terms, ensure you provide context to the viewer so the reason for posting is clear.").

557. *See* Steve Huffman, *Remember the Human—Black Lives Matter*, REDDIT BLOG (June 1, 2020), https://redditblog.com/2020/06/01/remember-the-human-black-lives-matter [https://perma.cc/X65X-BMDV].

with a specific provision outlining what is prohibited with examples.[558] The company also banned the subreddit "r/The_Donald," which was created by third parties (not related to Trump), for violating the hate speech policy.[559] Reddit explained:

> All communities on Reddit must abide by our content policy in good faith. We banned r/The_Donald because it has not done so, despite every opportunity. The community has consistently hosted and upvoted more rule-breaking content than average (Rule 1), antagonized us and other communities (Rules 2 and 8), and its mods have refused to meet our most basic expectations. . . . To be clear, views across the political spectrum are allowed on Reddit—but all communities must work within our policies and do so in good faith, without exception.[560]

Reddit is different from Twitter and YouTube in that Reddit's platform allows politically partisan groups to discriminate based on political viewpoint and party affiliation if the rules for the subreddit so stipulate. (Facebook also allows users to form political, social, or other groups with their own rules for the discussion.) However, if the subreddit's rules have no such restrictions, then it may be hard to ensure political bias does not creep into the content moderation decisions of user-moderators who are charged with a good deal of the responsibility in content moderation. For example, before the 2016 election, some criticized the moderators of the subreddit "r/politics"—a general political discussion group—for "leaning heavily in favor of Hillary Clinton."[561]

F. Snapchat

Starting in late 2017, Snapchat has offered an alternative form of social media—a section where users interact with their circle of friends ("social") and a different section called Discover where curated media partners of Snap offer content ("media") to users.[562]

---

558. *See Promoting Hate Based on Identity or Vulnerability*, REDDIT, https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/promoting-hate-based-identity-or [https://perma.cc/S3BS-4ZCW].

559. *See Update to Our Content Policy*, *supra* note 7.

560. *Id.*

561. *Make r/politics an Unbiased Subreddit for All Political Parties*, CHANGE.ORG, http://chng.it/6NWmVQQYT5 [https://perma.cc/DX7T-524D].

562. *See* Evan Spiegel, *How Snapchat Is Separating Social from Media*, AXIOS (Nov. 29, 2017), https://www.axios.com/how-snapchat-is-separating-social-from-media-1513307227-64cafea7-db16-4f30-ae8a-2891677d400b.html [https://perma.cc/G54J-CUDZ].

This bifurcated approach minimizes the spread of fake news and misinformation.[563] Snapchat's community standards succinctly explain what content is prohibited.[564] They do not provide details of the review process beyond a brief statement: "We review these reports to determine whether there is a violation of these Guidelines and any action needs to be taken."[565] The community guidelines also indicate that "[m]edia partners in Discover agree to additional guidelines, including the requirement that their content is accurate and where appropriate, fact-checked."[566] In the past, Snapchat's community standards recognized a goal of consistent enforcement: "We will do our best to enforce them consistently and fairly, and ultimately we'll try to do what we think is best in each situation, at our own discretion."[567] Snapchat removed this language in its September 2020 community guidelines, which do recognize a goal of applying the guidelines "to all Snapchatters, equally."[568] CEO Evan Spiegel stated that the company reviews political ads for misinformation.[569] Snapchat recognizes a newsworthiness exception to content moderation.[570] Snapchat's website does not appear to discuss an appeal of its enforcement decisions, but one Change.org petition indicated that Snapchat does not allow appeals.[571] It is unclear from Snap's

563. *See* Mike Shields, *Snap Suddenly Has a Leg up on Facebook and Google—but It Still Needs to Do 2 Things to Steal Their Advertisers*, BUS. INSIDER (Oct. 7, 2017, 8:47 AM), https://www.businessinsider.com/snapchats-closed-doors-keep-fake-news-out-2017-10 [https://perma.cc/5Q7Q-LXVN].

564. *See Community Guidelines*, SNAP INC., https://www.snap.com/en-US/community-guidelines [https://perma.cc/9HB7-DS99].

565. *Id.*

566. *Id.*

567. Melissa Chan, *Snapchat's New Guidelines Warn Sexting Teens: 'Keep Your Clothes On!,'* N.Y. DAILY NEWS (Feb. 27, 2015, 10:44 AM), https://www.nydailynews.com/news/national/snapchat-new-guidelines-warn-teens-clothes-article-1.2131544.

568. *See Community Guidelines, supra* note 564.

569. *See* Makena Kelly, *Snapchat CEO Says His Company Fact-Checks Political Ads, Unlike Facebook*, VERGE (Nov. 18, 2019, 1:38 PM), https://www.theverge.com/2019/11/18/20970958/snapchat-evan-spiegel-facebook-political-ads-fact-checks-election.

570. *See Community Guidelines*, *supra* note 564; *see also* Katie Benner, *Snapchat Discover Takes a Hard Line on Misleading and Explicit Images*, N.Y. TIMES (Jan. 23, 2017), https://www.nytimes.com/2017/01/23/technology/snapchat-discover-takes-a-hard-line-on-misleading-and-explicit-images.html.

571. *See Let Snapchat Users Appeal if Their Account Has Been Wrongfully Locked/Terminated*, CHANGE.ORG, https://www.change.org/p/snap-inc-let-snapchat-users-appeal-if-their-account-has-been-wrongfully-locked-terminated [https://perma.cc/6PCS-VMVV].

community standards how the company ensures its enforcement is consistent or nonpartisan, although the bifurcated approach to Snap's platform may itself reduce any concerns about political bias.

Snapchat has a feature called Discover, which provides users with a feed of content created by publishers selected by Snapchat.[572] Snapchat determines what content is promoted in Discover.[573] In June 2020, Snapchat decided not to allow Trump's account to display his content on Discover because Trump's comments (outside of Snapchat) about the George Floyd protests could be viewed as "incit[ing] racial violence and injustice."[574]

G. Twitch

Twitch is a social network in which users livestream themselves, often while playing video games.[575] Its community standards are listed on one page, with links to more extensive discussion of its standards for hateful conduct and harassment, sexual content, and music content.[576] Notably, Twitch may consider taking enforcement action for harassment that occurs off the site if it is directed at a Twitch user.[577] Twitch's community standards do not contain any specific policies for politicians or political ads. When it temporarily suspended the account of Trump for violating its standards for hateful conduct (related to his comments about Mexicans during a Tulsa rally in late June 2020[578]), a Twitch spokesperson explained that Twitch does not have a public interest exception, but applies the same approach to politicians as any other user: "Like anyone else,

572. *See* Josh Constine, *Snapchat Uncovers Discover*, TECHCRUNCH (June 7, 2016, 12:00 PM), https://techcrunch.com/2016/06/07/snapchat-discover-previews [https://perma.cc/PR74-4VE9].

573. *See* Jordan Wahl, *What Is Snapchat Discover: Fresh Content at Your Fingertips*, G2 (Oct. 10, 2018), https://learn.g2.com/snapchat-discover [https://perma.cc/7U3Y-DL4N].

574. Cecilia Kang & Kate Conger, *Snap Says It Will No Longer Promote Trump's Account*, N.Y. TIMES (June 3, 2020), https://www.nytimes.com/2020/06/03/technology/snapchat-trump.html.

575. *See About*, TWITCH, https://www.twitch.tv/p/en/about.

576. *See Community Guidelines*, TWITCH, https://www.twitch.tv/p/legal/community-guidelines [https://perma.cc/T2V5-4TUQ].

577. *Id.* ("We may take action against users for hateful conduct or harassment that occurs off Twitch services that is directed at Twitch users.").

578. *See* Kris Holt, *Twitch Restores Donald Trump's Account After a Two-Week Suspension*, ENGADGET (July 13, 2020), https://www.engadget.com/twitch-donald-trump-suspension-lifted-180048663.html [https://perma.cc/WM26-MWUQ].

politicians on Twitch must adhere to our Terms of Service and Community Guidelines . . . . We do not make exceptions for political or newsworthy content, and will take action on content reported to us that violates our rules."[579] Twitch was the only company surveyed that did not recognize a public interest exception. On December 9, 2020, Twitch announced an updated policy to combat hateful conduct, harassment, and sexual harassment on its platform, including some discussion of how Twitch determines violations.[580]

Twitch's community standards don't mention impartiality or nonpartisan enforcement, but they support "users who express diverse or unpopular points of view."[581] The policy regarding hateful conduct states that it "affords every user globally equal protections under this policy, regardless of their particular characteristics."[582] Twitch explains that the potential penalties for a violation include "removal of content, a strike on the account, and/or suspension."[583] Twitch allows the user to appeal a suspension of an account or a warning it issues.[584]

H. TikTok

TikTok is the latest craze in social media, enabling people to share short videos through an internet platform.[585] TikTok is different from the other platforms discussed because it originates from China—a source of controversy due to concerns of data collection.[586] Given its

---

579. Igor Bonifacic, *Twitch Has Suspended Donald Trump's Account*, ENGADGET (June 29, 2020), https://www.engadget.com/twitch-suspends-donald-trump-account-174145621.html [https://perma.cc/N3GF-ZSPK].

580. *See Introducing Our New Hateful Conduct & Harassment Policy*, TWITCH (Dec. 16, 2020), https://blog.twitch.tv/en/2020/12/09/introducing-our-new-hateful-conduct-harassment-policy [https://perma.cc/YPD8-V6AT].

581. *Hateful Conduct and Harassment [NEW]*, TWITCH, https://www.twitch.tv/p/legal/community-guidelines/harassment/20210122 [https://perma.cc/K8XG-NMPZ].

582. *Id.*

583. *Community Guidelines*, *supra* note 576.

584. *See About Account Enforcements and Chat Bans*, TWITCH, https://help.twitch.tv/s/article/about-account-suspensions-dmca-suspensions-and-chat-bans?language=en_US.

585. *See* Claire Pedersen et al., *Inside the TikTok Craze and Why There Are Concerns over Chinese Data Collection, Censorship*, ABC NEWS (Nov. 5, 2019, 8:30 PM), https://abcnews.go.com/Business/inside-tiktok-craze-concerns-chinese-data-collection-censorship/story?id=66768839 [https://perma.cc/GB7Y-F8EY].

586. *See id.* (noting that Tiktok is owned by ByteDance, a Chinese artificial intelligence company).

connection to China, President Trump had attempted to ban TikTok by invoking powers under the International Emergency Economic Powers Act[587] (IEEPA), but two federal judges granted temporary injunctions against the enforcement of the Secretary of Commerce's prohibitions on TikTok, which the courts held likely went beyond the authority provided by IEEPA.[588] It was unclear in January 2021 whether the Biden Administration would take a different approach to TikTok.

TikTok's community standards include, for each standard, a very helpful section instructing "Do not post," with examples of what users should not post.[589] Similar to Twitter, TikTok bans all political ads: "Ads must not reference, promote or oppose a candidate for public office, current or former political leader, political party, or political organization. They must not contain content that advocates . . . (for or against) [ ] a local, state, or federal issue of public importance."[590] But "[c]ause-based advertising or public service announcements from non-profits or government agencies may be allowed, if not driven by partisan political motives."[591] Like YouTube and Twitch, TikTok updated its community guidelines in December 2020; TikTok did so to foster well-being on the platform, for example, to address suicide, self-harm, and distressing content.[592] TikTok's policy against misinformation consisting of "[c]ontent that misleads community members about elections or other civic processes" remained the same.[593] According to one report, TikTok removed videos containing

587. 50 U.S.C. §§ 1701–07.

588. *See* TikTok Inc. v. Trump, No. 1:20-cv-02658-CJN, 2020 WL 7233557, at *15, *18 (D.D.C. Dec. 7, 2020); Marland v. Trump, No. 20-4597, 2020 WL 6381397, at *12, *15 (E.D. Pa. Oct. 30, 2020).

589. *See Community Guidelines*, TIKTOK, https://www.tiktok.com/community-guidelines?lang=en [https://perma.cc/QL67-RJMT].

590. *TikTok Advertising Policies—Ad Creatives*, TIKTOK, https://ads.tiktok.com/help/article?aid=9552 [https://perma.cc/UX9L-8RVU]; *see also* Blake Chandlee, *Understanding Our Policies Around Paid Ads*, TIKTOK, https://newsroom.tiktok.com/en-us/understanding-our-policies-around-paid-ads [https://perma.cc/J4VZ-AUDG].

591. *TikTok Advertising Policies—Ad Creatives*, *supra* note 590.

592. *See* Cormac Keenan, *Refreshing Our Policies to Support Community Well-Being*, TIKTOK (Dec. 15, 2020), https://newsroom.tiktok.com/en-us/refreshing-our-policies-to-support-community-well-being [https://perma.cc/TZZ7-A5JG]; *Community Guidelines*, TIKTOK (last updated Dec. 2020), https://www.tiktok.com/community-guidelines?lang=en [https://perma.cc/VL3A-RD8Q].

593. *See Community Guidelines*, *supra* note 592.

election misinformation, including QAnon conspiracy theories about ballots, but not before the videos amassed over 200,000 views.[594]

TikTok uses "a mix of technology and human moderation" and invites users to report violations.[595] Based on the H1 2020 Transparency Report, TikTok appears to rely heavily on filtering or moderation via technology.[596] Like most of the other internet platforms, TikTok's community standards do not expressly adopt a principle of nonpartisanship, although they "apply to everyone and to everything on TikTok."[597] TikTok recognizes a public interest exception allowing violating content to remain on TikTok "under certain circumstances, such as educational, documentary, scientific, or artistic content, satirical content, content in fictional settings, counterspeech, and content in the public interest that is newsworthy or otherwise enables individual expression on topics of social importance."[598]

TikTok has undertaken major initiatives to make their procedures more transparent—even beyond other Internet platforms' disclosures. TikTok announced the formation of a Transparency Center in Los Angeles to allow people to examine the company's content moderation.[599] This initiative comes amidst the United States and other countries' concerns about privacy and putative surveillance on TikTok.[600] In March 2020, TikTok announced a Content Advisory

---

594. *See* Kari Paul, *TikTok: False Posts About US Election Reach Hundreds of Thousands*, GUARDIAN (Nov. 5, 2020, 7:30 PM), https://www.theguardian.com/technology/2020/nov/05/tiktok-us-election-misinformation [https://perma.cc/NV2E-FZR8].

595. *Community Guidelines*, *supra* note 592.

596. *See* Michael Beckerman, *Our H1 2020 Transparency Report*, TIKTOK (Sept. 22, 2020), https://newsroom.tiktok.com/en-us/our-h-1-2020-transparency-report [https://perma.cc/6MUD-R8Q4] (noting TikTok "removed 96.4% of these videos before they were reported to us, and 90.3% were removed before they received any views").

597. *Community Guidelines*, *supra* note 592.

598. *Id.*

599. *See* Vanessa Pappas, *TikTok to Launch Transparency Center for Moderation and Data Practices*, TIKTOK (Mar. 11, 2020), https://newsroom.tiktok.com/en-us/tiktok-to-launch-transparency-center-for-moderation-and-data-practices [https://perma.cc/HK4E-2JJE]; Casey Newton, *Three Takeaways from a Visit to TikTok's New Transparency Center*, VERGE (Sept. 11, 2020, 6:00 AM), https://www.theverge.com/interface/2020/9/11/21430822/tiktok-transparency-visit-tour-algorithms-for-you-page.

600. *See* Christopher Brito, *U.S. "Looking at" Banning TikTok and Other Chinese Social Media Apps, Mike Pompeo Says*, CBS NEWS (July 7, 2020, 3:02 PM), https://www.cbsnews.com/news/tiktok-pompeo-united-states-weighing-ban-chinese-social-

Council of independent experts chaired by Professor Dawn Nunziato.[601] The Council will advise TikTok on "critical topics around platform integrity, including policies against misinformation and election interference."[602] These initiatives indicate that TikTok's content moderation is still a work-in-progress. As a company representative said on July 9, 2020, "We're working every day to be more transparent about the violating content we take down and offer our users meaningful ways to have more control over their experience, including the option to appeal if we get something wrong."[603] Then-CEO Kevin Mayer announced on July 29, 2020 that TikTok was launching a "Transparency and Accountability Center for moderation and data practices," premised on the belief that "all companies should disclose their algorithms, moderation policies, and data flows to regulators."[604] TikTok later disclosed its algorithms, a first for any internet platform.[605]

I. Internet Platforms' Internal (Nonpublic) Manuals

Internet platforms may have internal company manuals that set forth more detailed guidelines, including nonpartisanship or impartiality as a principle for their content moderators.[606] If so, they should publicize the internal standards that content moderators use, as well as the

---

media-apps [https://perma.cc/Z6S6-FXMM] (noting India has already banned TikTok and Australia is considering it).

601. *See* Vanessa Pappas, *Introducing the TikTok Content Advisory Council*, TIKTOK (Mar. 18, 2020), https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council [https://perma.cc/HRG7-CFLK].

602. *Id.*

603. Jonathan Chadwick, *TikTok Deleted Almost 50 MILLION Videos in Just Six Months and Received 500 Legal Requests for User Data from Governments Around the World*, DAILY MAIL (July 10, 2020, 6:49 AM), https://www.dailymail.co.uk/sciencetech/article-8507003/TikTok-removed-49-million-videos-six-months-breaking-content-rules.html [https://perma.cc/HBB4-7VNM].

604. Kevin Mayer, *Fair Competition and Transparency Benefits Us All*, TIKTOK (July 29, 2020), https://newsroom.tiktok.com/en-us/fair-competition-and-transparency-benefits-us-all [https://perma.cc/TV8L-DRF7].

605. *See* Sara Fischer, *Inside TikTok's Killer Algorithm*, AXIOS (Sept. 10, 2020), https://www.axios.com/inside-tiktoks-killer-algorithm-52454fb2-6bab-405d-a407-31954ac1cf16.html [https://perma.cc/868A-BU88]; Casey Newton, *TikTok Has a Bold New Plan to Win over Regulators*, VERGE (July 31, 2020, 6:00 AM), https://www.theverge.com/interface/2020/7/31/21348172/tiktok-algorithms-transparency-accountability-review-lawmakers-michael-beckerman-interview.

606. *See, e.g.*, Klonick, *supra* note 61, at 1633 (describing internal "booklet" for content moderation at YouTube).

exact procedures that are in place to promote nonpartisan content moderation of political candidates and political ads. The lack of transparency allows the allegations of "political bias" to fester. As the above survey shows, the internet platforms all fail to provide much, if any, information explaining how they ensure nonpartisan content moderation of political candidates.

III. THE CASE FOR NONPARTISANSHIP AS A COMMUNITY STANDARD FOR CONTENT MODERATION OF POLITICAL CANDIDATES AND POLITICAL ADS

Part III sets forth the affirmative case for why internet platforms should recognize nonpartisanship as a community standard that they follow in moderating the content of political candidates and public officials—and should establish transparent procedures to ensure adherence to this important principle.

A. Why Nonpartisanship in Content Moderation Matters

Large internet platforms that offer themselves as fora for wide open public discussion should adopt a principle of nonpartisanship in moderation of the content of political candidates. This principle is consistent with free and open exchange of speech, particularly on political issues related to elections.

1. Partisan versus nonpartisan content moderation of politicians' content

This Article does not explore the larger question whether internet platforms should be neutral or nonpartisan in general for all content moderation of the millions of users and the billions of content on their platforms. That question is left for future inquiry.

Instead, this Article focuses on the narrower issue involving the online content of elected officials and political candidates running for office in the United States, as well as political campaign ads by those candidates or political action committees. (Other government employees or non-elected officials are excluded.) The class of political candidates and public officials presents a more finite and limited number of people. If operationalizing a principle of nonpartisanship would require greater staff and resources, then it would be far more realistic for platforms to deal first with potentially 537 federal officials, 18,749 state officials, and 500,396 local officials,

just in the United States.[607] Though that figure is large, it pales in comparison to the 2.7 billion active users on Facebook.[608]

This Article defines the *principle of nonpartisanship* as the review of the content of a political candidate or a political campaign ad, for potential violations of the community standards of an internet platform, without bias or favoritism due to the political party of the candidate or the person or group who posted the campaign ad. Thus, if content moderators find a violation of the community standards because of the candidate's party affiliation, either for or against, that decision would violate the principle of nonpartisanship. Decisions of content moderation should be politically nonpartisan—meaning they should not be based on the moderator's allegiance to one political party or another, nor should they be based on deference or favoritism to the political party in power. The company should not modify its content moderation decisions or policies to curry favor or avoid reprisal from a political leader, either.

Imagine that the CEO of an internet platform is a supporter of Candidate A because the CEO believes Candidate A is better for the platform's business. But an inflammatory post by Candidate A was flagged by the company's regular content moderation procedure because it violated the company's community standard against hate speech. However, knowing the CEO's support of Candidate A, a high-level executive who participates in the final review by its content moderation group reversed the violation decision and let Candidate A's post remain unmoderated for all to see. In short, the company executive knowingly engaged in political favoritism to Candidate A during its content moderation because the CEO politically supports Candidate A. Problematic?

Yes. Under the principle of nonpartisanship, the executive made a content moderation decision based on the CEO's personal preference for a candidate and overrode the violation decision because of a preference for the candidate, as opposed to the actual content in the post. Whatever the company's motivation in preferring a political candidate (e.g., agreement with the candidate's views,

---

607. *How Many Politicians Are There in the USA?*, POLIENGINE, https://poliengine.com/blog/how-many-politicians-are-there-in-the-us [https://perma.cc/QWE8-QA6G].

608. J. Clement, *Facebook: Number of Monthly Active Facebook Users Worldwide 2008–2020*, STATISTA (Aug. 10, 2020), https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide.

belief the candidate is better for the company's business, fear the candidate will retaliate against the company if elected), the preference for the political candidate amounts to political support for the candidate—a factor that is wholly inappropriate to base a decision of content moderation. In short, the content moderation was partisan.[609]

2. Why a principle of nonpartisanship for political candidates should be recognized

a. Allowing voters' information from political candidates relevant to government and elections without political bias

Recognizing the proposed, limited principle of nonpartisanship should not be controversial. Except for Reddit, which allows politically partisan discussion groups to exclude comments from the opposing party consistent with the subreddit's rules, none of the internet platforms analyzed above suggest they allow their own content moderation to be partisan or politically biased. Twitter, Facebook, and YouTube openly tout their goal of consistent, fair, or uniform application of their community guidelines. The limited principle of nonpartisanship proposed by this Article is consistent with that general goal.

The principle of nonpartisanship is founded on the belief that voters benefit from having "political speech in the course of elections, the speech upon which democracy depends."[610] With such information, voters can evaluate the content from political candidates and political campaigns. Such content is critical to people's right to information about their government and their ability to make

609. Some critics question whether there's any evidence of this kind of partisan decision ever happening at social media companies. I am not privy to the internal decisions of internet platforms. However, media reporting has provided evidence of possible irregularities or political bias in content moderation decisions or policy. *See, e.g.*, *supra* notes 34–35, 318, 335; *infra* notes 697, 708–10 and accompanying text; Lauren Frayer, *Facebook Accused of Violating Its Hate Speech Policy in India*, NPR (Nov. 27, 2020, 3:46 PM), https://www.npr.org/2020/11/27/939532326/facebook-accused-of-violating-its-hate-speech-policy-in-india [https://perma.cc/QRS2-NZFS]; Whitney Tesi, *Facebook and Twitter Take Steps to Limit Spread of Controversial New York Post Article*, SLATE (Oct. 14, 2020, 6:31 PM), https://slate.com/technology/2020/10/hunter-biden-new-york-post-twitter-facebook-block.html [https://perma.cc/QDY7-FR5P] (reporting Evelyn Douek, a doctoral student at Harvard Law School, suggested that Twitter's decision on Hunter Biden story did not comport with its own policy on hacked materials).

610. Nixon v. Shrink Mo. Gov't PAC, 528 U.S. 377, 405 (2000) (Kennedy, J., dissenting).

informed decisions about voting.[611] This justification comports with the Supreme Court's general approach to scrutinizing federal disclosure requirements on campaign ads in federal elections.[612] As Richard Briffault summarizes, "By emphasizing the voter information that disclosure generates, disclosure actually 'further[s] First Amendment values by opening the basic processes of our federal election system to public view,'" under the Court's jurisprudence.[613] This approach is also consistent with Article 19(2) of the International Covenant on Civil and Political Rights, which recognizes: "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice."[614] The United States ratified this treaty in 1992.[615] Internet platforms can also benefit from incorporation of human rights standards in their content moderation policies, as David Kaye, the former UN special rapporteur on the promotion of the right to freedom of opinion and expression, has advocated.[616]

However, recognizing a principle of nonpartisanship does not mean that internet platforms cannot fact-check political ads or must allow election misinformation, hate speech, or attempts to suppress voters by posts or ads. To the contrary, given the lessons of election interference in the 2016 U.S. election, it would be Panglossian for internet platforms to sit back and allow rampant misinformation and foreign interference without moderation. Indeed, it would be gross negligence for internet platforms to allow their websites to be exploited—or weaponized—to suppress voters or to interfere with

---

611. *See* Carroll v. President & Comm'rs of Princess Anne, 393 U.S. 175, 182 (1968) ("It is vital to the operation of democratic government that the citizens have facts and ideas on important issues before them." (quoting A Quantity of Copies of Books v. Kansas, 378 U.S. 205, 224 (1964))).

612. *See* McConnell v. Fed. Election Comm'n, 540 U.S. 93, 196 (2003) (discussing the approach the Court enumerated in *Buckley v. Valeo*, 424 U.S. 1 (1976)).

613. Richard Briffault, *Two Challenges for Campaign Finance Disclosure After* Citizens United *and* Doe v. Reed, 19 WM. & MARY BILL RTS. J. 983, 991 (2011).

614. International Covenant on Civil and Political Rights art. 19(2), Dec. 19, 1966, 999 U.N.T.S. 171.

615. Jimmy Carter, *U.S. Finally Ratifies Human Rights Covenant*, CARTER CTR. (June 28, 1992), https://www.cartercenter.org/news/documents/doc1369.html [https://perma.cc/KZM9-K2E3].

616. *See* DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET 119 (2019).

elections. As long as the internet platform applies its community standards in nonpartisan fashion, it can moderate *everything*, including election misinformation. Nonpartisan treatment is all that is required.

Each internet platform can decide whether to exempt politicians from any of its community standards. Based on the survey in Part I, only Facebook did so by recognizing an exception for politicians and political ads from fact-checking, although Facebook altered its policy in September 2020, as noted above.[617] Facebook applies the same general community standards for hate speech, violence, and voter suppression to politicians and laypeople alike, as do the other internet platforms.[618] Moreover, all of the platforms except for Twitch recognize a "public interest" or "newsworthiness" exception to allow violating content to remain viewable, with a notation of the violation, on its site.[619] This labeling of violations provides a less restrictive alternative to removal of a politician's post, thereby balancing the interests of the general public in receiving the information and the interests of the internet platform in providing a safe forum for all of its users.[620]

b. Avoiding the filter bubble

Another reason why a principle of nonpartisanship should be adopted in content moderation is to ward off further entrenchment of filter bubbles among internet platforms. Internet theorist and activist Eli Pariser has warned about the dangers of how internet platforms feed information to their users based on algorithms that lack transparency and that may prioritize what the algorithm thinks the users like: "The danger of these filters is that you think you are getting a representative view of the world and you are really, really not, and you don't know it."[621] If content moderation itself becomes partisan, this problem will only worsen.

---

617. *See supra* Table 1.

618. *See Community Standards*, *supra* note 462.

619. *See supra* Table 1.

620. Some question the effectiveness of labels as a way to combat misinformation. *See* Brian Fung, *Social Media Bet on Labels to Combat Election Misinformation. Trump Proved It's Not Enough*, CNN (Dec. 8, 2020, 7:11 AM), https://www.cnn.com/2020/12/08/tech/facebook-twitter-election-labels-trump/index.html [https://perma.cc/XZ85-8PQU].

621. Jasper Jackson, *Eli Pariser: Activist Whose Filter Bubble Warnings Presaged Trump and Brexit*, GUARDIAN (Jan. 8, 2017, 8:00 AM), https://www.theguardian.com/media/2017/jan/08/eli-pariser-activist-whose-filter-bubble-warnings-presaged-trump-and-

Consider the controversy over Section 230. Republican lawmakers decried the content moderation of Trump by Twitter and other companies as censorship and election interference.[622] They also touted the internet platform Parler as a Twitter-alternative and an "unbiased" platform.[623] But Parler reportedly terminated the accounts of liberal users.[624] Instead of being a more open platform, Parler has become known as a conservative platform.[625] Whether accurate or not, the prospect of enclaves of conservative internet platforms and liberal ones should give us pause. It would effectively create internet fiefdoms or filter bubbles—exposing users to a one-sided feed of content. But, as Pariser warned: "If you only see posts from folks who are like you, you're going to be surprised when someone very unlike you wins the presidency."[626]

c. Revitalizing a commitment to common good over factions

Adopting a principle of nonpartisan content moderation can also be a way to revitalize a commitment to a common good in the United States. Internet platforms can effectively set the tone for a healthier debate by and over political candidates. Internet platforms set up basic ground rules for online debate in their community standards—and then moderate in a nonpartisan way.

Madison's famed *Federalist No. 10* sets forth a defense of a republican form of government in the Constitution as a way to counter the dangers of factions that inevitably arise.[627] One of the dangers of factions—what we might call today tribalism[628]—is that people become "divided [ ] into parties, inflamed [ ] with mutual animosity, and rendered [ ] much more disposed to vex and oppress

---

brexit [https://perma.cc/9DD5-PYHE]; *see* ELI PARISER, THE FILTER BUBBLE: WHAT THE INTERNET IS HIDING FROM YOU (2011).

622. *See* Fung et al., *supra* note 11.

623. *See* Danielle Abril, *Conservative Social Media Darling Parler Discovers that Free Speech Is Messy*, FORTUNE (July 1, 2020, 3:00 PM), https://fortune.com/2020/07/01/what-is-parler-conservative-free-speech-misinformation-hate-speech-john-matze [https://perma.cc/ANH4-UM2W].

624. Watts, *supra* note 12.

625. *See id.*

626. Jackson, *supra* note 621.

627. *See* THE FEDERALIST NO. 10 (James Madison).

628. *See* Joseph Russomanno, *Tribalism on Campus: Factions, iGen and the Threat to Free Speech*, 24 COMMC'N L. & POL'Y 539, 557 (2019).

each other than to co-operate for their common good."[629] In words that still ring true today, Madison identified the pervasiveness of factions: "A zeal for different opinions concerning religion, concerning government, and many other points, as well of speculation as of practice; an attachment to different leaders ambitiously contending for pre-eminence and power; or to persons of other descriptions whose fortunes have been interesting to the human passions . . . ."[630]

Madison's proposed solution was not to eliminate factions, which would require destroying liberty, but to control their pernicious effects, to make them "unable to concert and carry into effect schemes of oppression."[631] Madison believed the republican form of government set forth in the Constitution would provide such a check on factions.[632]

Whether Madison's view is persuasive is open to debate, especially given the rise of partisan politics.[633] Yet Madison's insight about the importance of designing institutions in a way to check oppression and to foster the ability of people to recognize "both the public good and the rights of other citizens" still holds true today.[634] Internet platforms that offer public forums for their users to engage in public discussion and debate should view themselves as architects of public, online spaces. The platforms should design institutional features that promote a common good and respect for all individuals instead of enabling factions to dominate or manipulate the conversation and engagement on a platform. Perhaps this idea sounds utopian. But if it is, our republic may be lost.

Just imagine if Facebook became Foxbook and catered only to conservative viewpoints or YouTube became CNNTube, the video sharing site for liberals. Extending partisanship to even more sectors, including social media, could be harmful to the functioning of democratic government, particularly in today's highly polarized climate. As Gallup senior scientist Frank Newport warns: "[P]olarization and partisan conflict lead to inaction, as 'my way or the highway,'

---

629. THE FEDERALIST NO. 10, at 59 (James Madison) (Jacob E. Cooke ed., 1961).

630. *Id.* at 58–59.

631. *Id.* at 61.

632. *See id.* at 60.

633. *See* Michael Gerhardt & Jeffrey Rosen, *How to Revive Madison's Constitution*, ATLANTIC (Dec. 4, 2019), https://www.theatlantic.com/ideas/archive/2019/12/madison-constitution/602929.

634. THE FEDERALIST NO. 10, *supra* note 629, at 60–61.

ideologically rigid mentalities lower the probability of achieving the compromise that should be at the heart of legislative functioning."[635]

d. Internet platforms for user-generated content are different from other media and publishers

The most common pushback I have received to my proposal is the argument that internet platforms like Facebook and Twitter should be treated no different from Fox News, CNN, book publishers, and other media, who are under no obligation to be nonpartisan even during elections. By this reasoning, Facebook or Twitter should be allowed to favor the content of a particular politician because it wants that politician to win. For example, one can point to television news networks (e.g., Fox News, CNN, and MSNBC) and newspapers (e.g., *Washington Times*, *Wall Street Journal*, and *New York Times*) as being partisan, to some extent, in presenting the news more favorably to conservative or liberal views.[636]

What this argument ignores is that internet platforms are different from newspapers, TV networks, and book publishers. As an initial matter, internet platforms are the only ones eligible to qualify for Section 230 immunity—which at the very least suggests that Congress viewed them differently. More importantly, internet platforms are open fora to potentially millions, if not billions, of users. Internet platforms reach a much greater audience—Facebook has 190 million users in the United States and Twitter, 68.7 million, while the *New York Times* has six million subscribers and Fox News around four million viewers.[637] Internet platforms also invite their millions of users

635. Frank Newport, *The Impact of Increased Political Polarization*, GALLUP (Dec. 5, 2019), https://news.gallup.com/opinion/polling-matters/268982/impact-increased-political-polarization.aspx [https://perma.cc/D6D3-EMWU].

636. *See AllSides Media Bias Chart*, ALLSIDES, https://www.allsides.com/media-bias/media-bias-chart [https://perma.cc/LRK7-U33F] (showing media sources' biases on a scale from progressive to conservative).

637. J. Clement, *Leading Countries Based on Facebook Audience Size as of October 2020*, STATISTA (Nov. 24, 2020), https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users; J. Clement, *Leading Countries Based on Number of Twitter Users as of October 2020*, STATISTA (Oct. 29, 2020), https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries; Sarah Scire, *The New York Times' Success with Digital Subscriptions Is Accelerating, Not Slowing down*, NIEMANLAB (May 6, 2020, 4:02 PM), https://www.niemanlab.org/2020/05/the-new-york-times-success-with-digital-subscriptions-is-accelerating-not-slowing-down [https://perma.cc/9V5W-GSDW]; Joseph Wulfsohn, *Fox News Reaches Highest Viewership in Network's History, Topping MSNBC, CNN in 2020*,

to publish user-generated content—typically without any condition that the content must serve a conservative or liberal viewpoint to qualify for publication. Internet platforms operate as a public space for their users to exchange information.[638] This important feature is what makes the media "social."[639] By contrast, television networks, newspapers, and book publishers select everything they publish. None of the television networks, newspapers, or book publishers offer people a platform for user-generated content. They are fundamentally different from internet platforms.

Internet platforms should operate in a way analogous to universities when it comes to political candidates. Under section 501(c)(3) of the Treasury Department regulations for tax-exempt nonprofits, universities lose their tax-exempt status if they "participat[e] or interven[e] in a political campaign on behalf of or in opposition to a candidate includ[ing] . . . the publication or distribution of written or printed statements or the making of oral statements on behalf of or in opposition to such a candidate."[640] This restriction (also known as the Johnson amendment for then-Senator Lyndon Johnson, who sponsored it) is designed to avoid universities taking on a partisan role while educating students, who may be impressionable and also a captive audience.[641] Universities have traditionally recognized academic freedom and diversity of viewpoints among faculty and students.[642] Allowing universities to stake out partisan positions related to elections would undermine these values. Yet, academic freedom does not mean that universities cannot have campus regulations against hate speech and racist comments, such as in the classroom or dorms.[643] Similarly, internet platforms have traditionally recognized freedom of expression and diversity of viewpoints as important parts of their mission, while also moderating hate

---

FOX NEWS (Feb. 25, 2020), https://www.foxnews.com/media/highest-viewership-network-history-msnbc-cnn-2020 [https://perma.cc/YPC8-725X].

638. *See* GILLESPIE, *supra* note 1, at 18–19.

639. *See id.* at 16–17.

640. Treas. Reg. § 1.501(c)(3)(iii) (2019); *see* 26 U.S.C. § 501(c)(3).

641. *See generally* Michael Fresco, Note, *Getting to "Exempt!": Putting the Rubber Stamp on Section 501(c)(3)'s Political Activity Prohibition*, 80 FORDHAM L. REV. 3015, 3019–21 (2012).

642. *See* David M. Rabban, *A Functional Analysis of "Individual" and "Institutional" Academic Freedom Under the First Amendment*, 53 L. & CONTEMP. PROBS. 227, 232–33 (1990).

643. *See* Melissa Weberman, Note, *University Hate Speech Policies and the Captive Audience Doctrine*, 36 OHIO N.U. L. REV. 553, 575–80 (2010).

speech.[644] Although all internet platforms must moderate illegal and harmful content, engaging in partisan content moderation of political candidates undermines the overarching values of free expression in a democracy.

In the end, it boils down to this: do we really want a country in which Zuckerberg, Dorsey, or other social media executives can steer their platforms to favor the political candidate of their choosing in a partisan manner—and influence the outcome of an election?

3. What if content that violates community standard aligns with a political party's positions?

It is important to distinguish partisan content moderation of a political candidate from the general application of the community standards in nonpartisan manner. If an internet platform has a community standard and applies it equally to political candidates regardless of party affiliation, then it is permissible under the principle of nonpartisanship. And if a person repeatedly violates the community standards, the person's account, whether political candidate or not, can be suspended according to the company's policy.[645] The principle of nonpartisanship proposed by this Article is not as broad as "neutrality" touted in some Section 230 reforms.[646] Some of these reforms want to require as a condition of Section 230 that internet platforms must be viewpoint neutral.[647]

The *Wall Street Journal*'s editors identified the problem with this overbroad viewpoint neutrality—it would gut the whole notion of content moderation.[648] As the editors convincingly wrote:

> Do Islamism and white separatism count as "political viewpoints," in which case muting extremists could be counted as "bias"? Could a site be dinged for booting Louis Farrakhan or Alex Jones, the conspiracist who has called the Sandy Hook shooting a hoax? Maybe the courts would be asked to sort it out. The Constitution protects fringe views, but it doesn't require Twitter or Facebook to disseminate them.

---

644. *See, e.g.*, *Hateful Conduct Policy*, TWITTER, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [https://perma.cc/QH9R-TNG3].

645. *See, e.g.*, *Twitter Bans Account of Former KKK Leader David Duke*, REUTERS (July 31, 2020, 5:15 PM), https://www.reuters.com/article/us-twitter-david-duke/twitter-bans-account-of-former-kkk-leader-david-duke-idUSKCN24W2CD.

646. *See supra* note 54 and accompanying text.

647. *See supra* Section I.C.2.b.

648. *See The Twitter Fairness Doctrine*, *supra* note 55.

> Section 230 was written to empower moderation, to keep the web from becoming a cesspool. Every minute, 500 hours of video are added to YouTube. Every day, Twitter gets something like 500 million tweets and Facebook one billion posts. In the world's worst game of Whac-A-Mole, their systems clobber untold quantities of jihadist propaganda, "revenge porn," snuff videos, attempts to "dox" enemies, and so on.[649]

Indeed, internet platforms perform a necessary screening of endless streams of offensive, dangerous, and illegal content. As the *Wall Street Journal* noted, the First Amendment protects some of this content; if the government were moderating propaganda by suspected terrorist groups or conspiracy theories about school shootings, it would likely violate the First Amendment as impermissible viewpoint discrimination.[650]

The Supreme Court's treatment of neutral laws of general applicability under the Free Exercise Clause is instructive. The Court has recognized "the general proposition that a law that is neutral and of general applicability need not be justified by a compelling governmental interest even if the law has the incidental effect of burdening a particular religious practice."[651] The Constitution does not require the government to give special treatment to an individual, invoking religion, to avoid a "valid and neutral law of general applicability on the ground that the law proscribes (or prescribes) conduct that his religion prescribes (or proscribes)."[652] By analogy, as long as an internet platform enforces its generally applicable community standards in nonpartisan fashion, it is permissible under the approach outlined here.

B. Why Best Practices Are Better Than Bills to Reform Section 230

1. Avoiding government entanglement in speech codes

A reason to prefer internet platforms to self-regulate using best practices is that the alternative of government regulation of online speech is a cure worse than the disease. The solution isn't for the government to start micromanaging content moderation policies or to impose heavy-handed speech codes, such as viewpoint neutrality,

649. *Id.*
650. *See id.*
651. Church of the Lukumi Babalu Aye, Inc. v. City of Hialeah, 508 U.S. 520, 531 (1993).
652. Emp. Div., Dep't of Hum. Res. of Or. v. Smith, 494 U.S. 872, 879 (1990) (quoting United States v. Lee, 455 U.S. 252, 263 n.3 (1982)).

on internet companies. Indeed, in repealing the FCC's net neutrality rules, FCC Chairman Ajit Pai, appointed by President Trump, defended the U.S.'s general policy of avoiding intrusive government regulation of the internet:

> President Clinton got it right in 1996 when he established a free market-based approach to this new thing called the [i]nternet, and the [i]nternet economy we have is a result of his light-touch regulatory vision . . . . We saw companies like Facebook and Amazon and Google become global powerhouses precisely because we had light-touch rules that apply to this [i]nternet. And the [i]nternet wasn't broken in 2015 when these heavy-handed regulations were adopted.[653]

It is important to recognize that Congress could not directly require internet platforms to adopt political viewpoint neutrality (or otherwise to limit their content moderation to only unlawful content) because such a law would violate the internet platforms' own freedom of speech.[654] In an analogous context, district courts have recognized that Google has a First Amendment right to render its search results as editorial judgments and opinions.[655] As Eugene Volokh and Donald Falk explain:

> A speaker is thus entitled to choose to present only the speech that "in [its] eyes comports with what merits" inclusion. And this right to choose what to include and what to exclude logically covers the right of the speaker to choose what to include on its front page, or in any particular place on that page.[656]

Under *Miami Herald Publishing Co. v. Tornillo*,[657] "the freedom to speak necessarily includes the right to choose what to include in one's

---

653. *See* Laurel Wamsley, *FCC's Pai: 'Heavy-Handed' Net Neutrality Rules Are Stifling the Internet*, NPR (Nov. 22, 2017, 12:10 PM), https://www.npr.org/sections/thetwo-way/2017/11/22/565962178/fccs-pai-heavy-handed-net-neutrality-rules-are-stifling-the-internet [https://perma.cc/3V64-N5P3].

654. *See supra* notes 74–75 and accompanying text.

655. *See, e.g.*, Zhang v. Baidu.com Inc., 10 F. Supp. 3d 433, 435 (S.D.N.Y. 2014); Langdon v. Google, Inc., 474 F. Supp. 2d 622, 630 (D. Del. 2007); Kinderstart.com, LLC v. Google, Inc., No. C06–2057, 2007 WL 831806, at *16 (N.D. Cal. Mar. 16, 2007); Search King, Inc. v. Google Tech., Inc., No. CIV-02-1457-M, 2003 WL 21464568, at *4 (W.D. Okla. May 27, 2003).

656. Eugene Volokh & Donald M. Falk, *Google: First Amendment Protection for Search Engine Search Results*, 8 J.L. ECON. & POL'Y 883, 887 (2012) (alteration in original) (footnote omitted).

657. 418 U.S. 241 (1974).

speech and *what to exclude*."[658] Likewise, Congress cannot compel internet companies to publish user content they disagree with.

Some of the proposed bills to amend Section 230 try to avoid this First Amendment problem by making the requirement of political viewpoint neutrality a precondition to qualify for the civil immunity afforded to internet platforms under Section 230.[659] This conditioning of a federal immunity from civil lawsuits (predominantly in state courts) implicates the Supreme Court's unconstitutional conditions doctrine.[660] Sometimes, the Court has described this doctrine broadly: the "overarching principle, known as the unconstitutional conditions doctrine, [ ] vindicates the Constitution's enumerated rights by preventing the government from coercing people into giving them up."[661] The government "may not deny a benefit to a person on a basis that infringes his constitutionally protected interests—especially, his interest in freedom of speech."[662] However, other times, the Court has viewed this doctrine narrowly, if not inconsistently.[663]

Even assuming Congress has the authority to condition Section 230 immunity on internet platforms' refraining from viewpoint discrimination, it would not be wise policy. Some of the proposed Section 230 bills would entangle the FTC or courts in thorny determinations of whether an internet platform engaged in content moderation in a "politically biased manner," beyond "unlawful material," not "in a viewpoint-neutral manner," or not justified by "a compelling reason for restricting that access or availability."[664] For example, the Ending Support for Internet Censorship Act would require internet platforms to obtain an "immunity certification from the [FTC]" as a prerequisite for the platforms to obtain civil immunity under Section 230.[665] Under the bill, the internet platforms would have to prove, by clear and convincing evidence,

---

658. Volokh & Falk, *supra* note 656, at 887 (emphasis added) (citing *Miami Herald Publ'g Co.*, 418 U.S. at 258).

659. *See supra* note 343 and accompanying text.

660. *See* Koontz v. St. Johns River Water Mgmt. Dist., 570 U.S. 595, 604 (2013) (explaining the unconstitutional conditions doctrine).

661. *Id.*

662. Perry v. Sindermann, 408 U.S. 593, 597 (1972).

663. *See* Dolan v. City of Tigard, 512 U.S. 374, 407 n.12 (1994) (Stevens, J., dissenting) ("[T]he 'unconstitutional conditions' doctrine has for just as long suffered from notoriously inconsistent application; it has never been an overarching principle of constitutional law that operates with equal force regardless of the nature of the rights and powers in question.").

664. *See supra* notes 355, 361, 367 and accompanying text.

665. Ending Support for Internet Censorship Act, S. 1914, 116th Cong. § 2 (2019).

"that the provider does not (and, during the 2-year period preceding the date on which the provider submits the application for certification, did not) moderate information provided by other information content providers in a politically biased manner."[666] The definition of "politically biased moderation" is broad. It covers both discriminatory intent and disparate impact (even without intentional discrimination).[667] Thus, if an internet platform's community standards against hate speech resulted in "disproportionately restrict[ing] . . . access to, or the availability of, information from a political party, political candidate, or political viewpoint," that result presumably would violate the requirement.[668] Under either type of claim, one can easily imagine that the FTC would have to sift through millions of posts on the platform from a two-year period.

Compare this expansive level of mandatory review by the FTC for every internet platform seeking Section 230 immunity (which would be all of them), with the Department of the Treasury (Treasury) regulation requiring universities to avoid political campaign activity for a political candidate for public office.[669] The Treasury regulation is limited to political campaigns; it does not prohibit political viewpoints, such as a university publicly supporting Black Lives Matter.[670] Moreover, unlike the Ending Support for Internet Censorship Act, the Treasury regulation does not require universities to prove, by clear and convincing evidence (or at all), that they have not engaged in political campaign activity for a political candidate for public office.[671] The Treasury Department does not examine most 501(c)(3) organizations for compliance.[672] And it is far easier for

666. *Id.*

667. *See id.*

668. *See id.*

669. *See* Treas. Reg. § 1.501(c)(3) (2019).

670. *See, e.g.*, *Making a Better World Together*, HARV. COLL., https://college.harvard.edu/about/deans-messages/making-better-world-together [https://perma.cc/9VLZ-R4KP] (statements of Danoff Dean of Harvard College Rakesh Khurana and Dean of Students Katie O'Dair).

671. *See supra* notes 355–56 and accompanying text.

672. *See* Staff of the Joint Comm. on Tax'n, 108th Cong., Description of Present Law Relating to Charitable and Other Exempt Organizations and Statistical Information Regarding Growth and Oversight of the Tax-Exempt Sector 37 (Comm. Print 2004) ("The number of exempt organization returns examined by the IRS declined from 12,589 returns in 1993 to 5,754 returns in 2003. During the period 1993 through 2003, the number of returns examined as a percentage of the number of returns filed declined from 2.5 percent to 0.7 percent.").

universities to regulate the speech of their employees to abide by the Treasury regulation than it would be for an internet platform, which must engage in content moderation of millions of posts by millions of users. It might be difficult, if not impossible, for an internet platform to engage in content moderation—such as removing coordinated attacks by bots and content from suspected terrorist groups or white supremacists—without exposing itself to the potential legal claims of viewpoint discrimination under some of the proposed Section 230 bills.[673] Indeed, the bills could open the floodgates to lawsuits against internet platforms. A court would then have to scrutinize whether the content that was moderated was politically neutral—a difficult issue to determine.

2. Internet platforms' need for flexibility

One danger in legislating the parameters of permissible content moderation for internet platforms as a prerequisite to Section 230 immunity is locking in a procrustean approach to the constantly evolving internet.[674] Limiting internet platforms to content moderation of only unlawful content, for example, would hamper the ability of internet platforms' ability to moderate legal content that is harmful or offensive, such as misinformation about COVID-19 that could result in greater deaths of people who believed the misinformation spread on the internet platforms. Likewise, eliminating the "otherwise objectionable" catchall in the current Section 230 would remove an important flexible standard that allows internet platforms to adjust to changing circumstances. Indeed, one of the stated policies of Section 230 is "to preserve the vibrant and competitive free market that presently exists for the [i]nternet and other interactive computer services, unfettered by Federal or State regulation."[675] Another advantage of flexibility is that internet companies can restrict some harmful speech—such as deepfake sex videos that depict real people in simulated sex without their consent—that Congress might not be able to proscribe without violating the First Amendment.[676]

---

673. *See supra* notes 352–56, 364–66, 392–93 and accompanying text.

674. *See* Edward Lee, *Rules and Standards for Cyberspace*, 77 NOTRE DAME L. REV. 1275, 1280 (2002) (describing courts' discussions of the rapidly changing nature of the internet).

675. 47 U.S.C. § 230(b)(2).

676. *See, e.g.*, Cass R. Sunstein, *Falsehoods and the First Amendment*, 33 HARV. J.L. & TECH. 387, 418–23 (2020) (proposing how the government may regulate or ban deepfakes consistent with the First Amendment); Mary Anne Franks & Ari Ezra

IV. MODEL FRAMEWORK FOR NONPARTISAN CONTENT MODERATION (NCM) OF POLITICAL CANDIDATES

This Part outlines a model framework for nonpartisan content moderation (NCM) of political candidates, elected public officials, and political ads. The model framework is not intended as the exclusive way internet platforms should handle this issue, much less as the panacea for concerns of political bias in content moderation. Instead, the model NCM framework is offered with the expectation that greater input and deliberation, as well as tailoring for each platform, will be necessary. It also must be underscored that the proposal does not apply to any content beyond political candidates, elected public officials, and political ads.

A. The Model NCM Framework

The model NCM framework has the overriding goals of politically nonpartisan content moderation and transparency. To achieve these goals, the NCM framework incorporates an array of features, including (1) a clear statement of the policy and the steps of enforcement, (2) the institution of three levels of double-blind review, which can be streamlined as needed for expedited review, (3) the inclusion of both a public advocate and civil rights advocate in the appeals process, (4) the recognition of a defense of selective enforcement afforded to the politician whose content is at issue, (5) the inclusion of a specific category of content moderation of politicians in transparency reports, and (6) the separation of powers that excludes the company's business and lobbying executives, including the CEO, from any involvement in the content moderation decisions. These features are explained below.

1. Internet platforms' clear statement of NCM policy

Transparency is the starting point. As shown in Part II, internet platforms have disclosed little about the specific procedures they use for content moderation, whether for politicians or regular users, to decide whether a violation has occurred.[677] Many scholars have roundly criticized the opaqueness of content moderation procedures

---

Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD. L. REV. 892, 897 (2019) (same).

677. *See supra* Section II.A (illustrating the limited disclosure of internet platforms and lack of detail in community standards posted online).

and decisions.[678] Perhaps internet platforms hope to keep the "underbelly" of their operations from public scrutiny, including the fact that some companies outsource a good portion of their content moderation overseas to contract workers.[679] This lack of transparency may be changing. As discussed in Section II.B above, Twitter has disclosed more of its content moderation than perhaps any other company, especially for civic integrity and its public interest exception, including a general sketch of the review involved for exercise of the public interest exception.[680] Facebook has likewise publicized its grand plan to establish an independent Oversight Board, funded by a separate trust, that will decide appeals of content removal or "significant and difficult cases" referred by Facebook.[681] The Oversight Board launched its own website outlining the appeals process.[682] Article I of its charter states that the Board members "will exercise neutral, independent judgment and render decisions impartially,"[683] and its code of conduct sets forth rules of independence, impartiality, and the avoidance of conflicts of interest.[684] Yet neither Facebook nor Twitter have disclosed the specific stages of their own internal processes for determining violations of their community standards.

All internet platforms should publish—on a dedicated page in their community standards—a clear statement on how they treat possible violations of the standards by political candidates and elected public officials, as well as in political campaign ads. The internet platforms should express their commitment to *nonpartisan* content moderation of political candidates. The policy should start out with an introduction explaining the goals, such as the following:

---

678. *See, e.g.*, KAYE, *supra* note 616, at 51; GILLESPIE, *supra* note 1, at 115–16; Klonick, *supra* note 61, at 1665.

679. *See* ROBERTS, *supra* note 420, at 105; Adrian Chen, *The Laborers Who Keep Dick Pics and Beheadings out of Your Facebook Feed*, WIRED (Oct. 23, 2014, 6:30 AM), https://www.wired.com/2014/10/content-moderation [https://perma.cc/UPZ9-BRTW].

680. *See supra* Section II.B (discussing Twitter's extensive community standards policy and commitment to civic integrity).

681. *See* Brent Harris, *Preparing the Way Forward for Facebook's Oversight Board*, FACEBOOK (Jan. 28, 2020), https://about.fb.com/news/2020/01/facebooks-oversight-board [https://perma.cc/7Z53-HSR5].

682. *Appealing Content Decisions on Facebook or Instagram*, OVERSIGHT BOARD, https://www.oversightboard.com/appeals-process [https://perma.cc/43VW-RXHD].

683. *Governance: Oversight Board Charter*, OVERSIGHT BOARD, https://www.oversightboard.com/governance [https://perma.cc/VDX3-XP4U].

684. OVERSIGHT BOARD, OVERSIGHT BOARD BYLAWS: CODE OF CONDUCT 38–39 (2020), https://www.oversightboard.com/sr/governance/bylaws [https://perma.cc/TX6K-ZK7T].

> OUR COMMITMENT TO NONPARTISAN CONTENT MODERATION OF POLITICAL CANDIDATES, ELECTED PUBLIC OFFICIALS, AND POLITICAL CAMPAIGN ADS
>
> Elections are vital to democracy. As an internet platform, we understand our platform facilitates political debate among people of all persuasions. We undertake this important responsibility with the utmost concern. At the same time, malicious efforts to interfere with the elections through misinformation, coordinated inauthentic behavior, and voter suppression are a real and constant threat to democracy. That's one important reason we moderate content users post on our site. All users, including politicians and elected public officials, are subject to the same community standards for misinformation, hate speech, coordinated inauthentic behavior, and other harmful content. No user is exempt from or above our community standards.
>
> We recognize, however, that, to make informed decisions, voters benefit from having more information from the candidates, not less. For that reason, we have additional procedures when handling possible violations of our community standards by political candidates and elected officials, or in campaign ads. Using these procedures, we still apply the same community standards of what constitutes a violation, but, if we find a violation, we may apply different remedial actions to allow some offending content by a political candidate or campaign ad to remain on our site that we would have removed had it been posted by a regular user. We recognize a "public interest" exception by which we determine if the public's interest in receiving the content outweighs its potential harm to our community. But we may moderate the violating content with a label or other restrictions. These procedures are designed to ensure our content moderation is nonpartisan and allows free exchange of ideas by political candidates so voters can decide, while at the same time preventing foreign interference, voter suppression, and election misinformation.

This statement begins by recognizing the company's understanding of its important role in allowing political debate related to elections on its platform and its responsibility to protect its community from election interference and violations of its community standards, including misinformation and hate speech. Following the lessons of Russian interference in the 2016 U.S. election, internet companies should take greater responsibility for how their platforms may be exploited in ways that may undermine elections. Of course, internet platforms can decide different goals. For example, some companies (e.g., Twitter, Reddit, TikTok) decided to restrict political ads or

other political content not consistent with their mission or standards.[685] Whatever the policy, the internet platform should provide a clear statement on how it treats content of political candidates, public officials, and campaign ads that potentially violate their community standards.

2. Clear statement of the NCM procedure for violations, penalties, and appeals

Once the company's goals have been clearly articulated, the statement should explain, step by step, the procedures for how it handles content moderation of political candidates, elected public officials, and campaign ads. It is a tall, if not impossible, task to devise a process that eliminates all possibility of political bias, including implicit bias. Yet courts and agencies routinely make decisions affecting people with the expectation that those decisions are not being made based on the political affiliation of the person involved.

What are the mechanisms to reduce partisanship in content moderation? This Article relies on several components: (1) the selection and training of moderators; (2) the verification and training of political candidates, public officials, and entities purchasing campaign ads; (3) three levels of double-blind review of content, which can be streamlined for expedited reviews; (4) multiple moderators at potentially three levels—initial review, panel review, and appellate board review—all of whom are randomly assigned to review a potential violation and some of whom are independent moderators outside the company; (5) automatic right of appeal if a violation is found; (6) a defense of selective enforcement afforded to the person whose content has been flagged; and (7) automatic appointment of a public advocate for each appeal and appointment of a civil rights advocate in cases involving hate speech, voter suppression, or other civil rights issues. A proposed model procedure is set forth below:

> NONPARTISAN CONTENT MODERATION (NCM) PROCEDURE FOR DETERMINING VIOLATIONS, PENALTIES, AND APPEALS
>
> [1] Our community standards set forth the rules all users must abide by. Users have a responsibility to learn and follow these standards.

685. *See supra* notes 439, 554, 591 and accompanying text.

[2] Candidates for political office, elected public officials, and entities promoting a political campaign ad can register with us. They must pass a screening to verify themselves and, if verified, their accounts will have a designation (P) as a political candidate or (PAC) as a political action committee. Such qualification entitles them to the following special procedures for content moderation.

[3] If the content of a verified political account (P) or (PAC) is flagged for possible violation of our community standards, the content is subject to a special system of review by different, randomly assigned moderators to ensure the review is conducted in a nonpartisan manner. The moderators do not know the identity of the user. The content is anonymized during the entire review.

[4] *First level of review*. In the first level of review, three moderators individually review the content and decide whether it violates our community standards. The moderators do not know what any other reviewer has decided. If the first level of review is unanimous in finding a violation, the panel agrees upon the penalty.

[5] *Second level of review*. If the first level of review does not result in a unanimous decision, a second level of review is conducted by a panel of three randomly selected moderators who meet and deliberate about the potential violation and reach a unanimous panel decision.

[6] *Penalties*. If a violation is found, the panel decides what penalty should follow. The penalties include (i) removal of the content, (ii) leaving the content online but with a label noting the violation, (iii) adding a warning screen that a viewer must click to view the content, (iv) downgrading the search ranking of the content on the platform, (v) quarantining the content so it cannot be shared on the platform, and (vi) suspension of the user account due to repeated violations. For the ultimate penalty of account suspension, we generally apply a "3 strikes and you're out" approach. A user found to have violated the community standards on three separate occasions, within the past three years, faces automatic, permanent suspension of the account. One year following such suspension, the user may apply for reinstatement upon the user's acceptance of responsibility for the past violations and commitment to abide by the community standards.

[7] *Public interest exception*. Under the public interest exception, the panel has the discretion to decide to leave the content online after weighing the public's interest in viewing the content versus the potential harm in further dissemination. If the panel decides the public interest outweighs, it will provide an explanation that will be included in a notation as a label next to the content. The notation will indicate that the user has violated our community standards.

> [8] *Right of appeal and third level of review.* Verified political accounts (P) and (PAC) can appeal a violation decision to our appeals board. For each appeal, a public advocate is appointed and participates as amicus curiae to present the interests of the public. If the content involves hate speech, voter suppression, or other civil rights issue, a civil rights advocate is also appointed to present the issues from a civil rights perspective.
>
> [9] *Final decision.* The appeals board will decide the appeal on the written submissions and publish a decision explaining its reasons to the user. During the appeal, the penalty will remain in effect, but if the content is still online, a further notation of the appeal will be added. If the appeals board upholds the decision, a further notation will be added with a link to the decision for the public to view.

The first two paragraphs above provide the background to the treatment of political candidates. It begins by stating clearly that every user is held to the community standards, with a link to the standards. Ideally, the company's explanations of the standards would include examples of violations. The second paragraph above then indicates that candidates who seek the special procedures for content moderation must be screened and verified by the company. Each candidate or group must watch a short training video prepared by the company explaining the community standards and the process for political candidates. Each candidate or group must pass a short quiz on the community standards contained in the training video, although it may be taken as many times as necessary to pass. If the user is approved, a "(P)" or "(PAC)" will appear on the user's profile on the internet platform, entitling the candidate or group to the special procedures.

3. Selection and training of moderators

The selection of moderators is important. Internet companies should use a combination of employees and outside moderators, all of whom are professionally trained in content moderation. For example, both Google and Facebook have decided to incorporate outside experts—in Google's administration of right to be forgotten claims and in Facebook's content removal subject to the Oversight Board—in some of their review.[686] Thus far, content moderation has been a murky field with a good deal of it outsourced to contract workers, but a portion of it is still reserved for a more professional cadre of moderators.[687] For

686. *See* Lee, *supra* note 67, at 1066–72; Harris, *supra* note 681.

687. *See* ROBERTS, *supra* note 420, at 44–47 (describing the various employment conditions for most content moderation workers).

the content moderation of political candidates, the review should not be outsourced. The review should be assigned to professionals who are well-trained in the issues and who are subject to ethical responsibilities of nonpartisanship and impartiality. Indeed, as a matter of best practices, moderators should be asked to sign a pledge to abide by the ethical responsibilities the company or an outside professional group has identified for content moderation. A promising development in this area is the recent launch of two new organizations, the Trust & Safety Professional Association (TSPA) and the Trust & Safety Foundation Project, to "support[] the global community of professionals who develop and enforce principles and policies that define acceptable behavior and content online."[688] It is also important that the moderators represent a diverse cross-section of society with different areas of relevant expertise and experience, but some common training in the values the company seeks to protect (e.g., free speech, civil rights, free and fair elections, anti-harassment, impartiality).[689] Diversity of workforce can help to curb implicit biases.[690]

4. Three levels of randomized double-blind review, and inclusion of public advocate and civil rights advocate

The model framework is based on three levels of randomized double-blind review. Double-blind review means that neither the user nor the moderator will know the identity of each other.[691] Randomized double-blind review is a cornerstone of review of clinical trials and is meant to eliminate the bias that can result if the identity is known.[692] Random assignment of moderators provides a check against selection

688. *About*, TR. & SAFETY PRO. ASS'N, https://www.tspa.info [https://perma.cc/2DCC-6E47]; *see About*, TR. & SAFETY FOUND. PROJECT, https://www.tsf.foundation [https://perma.cc/5NWM-KU6R].

689. *See* GILLESPIE, *supra* note 1, at 201–02.

690. *See* Lori Mackenzie & Shelley J. Correll, *Two Powerful Ways Managers Can Curb Implicit Biases*, HARV. BUS. REV. (Oct. 1, 2018), https://hbr.org/2018/10/two-powerful-ways-managers-can-curb-implicit-biases; Vivian Hunt et al., *Delivering Through Diversity*, MCKINSEY & CO. (Jan. 18, 2018), https://www.mckinsey.com/business-functions/organization/our-insights/delivering-through-diversity [https://perma.cc/VT8J-EEFG].

691. *See* Adrian Mulligan et al., *Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers*, 64 J. AM. SOC'Y FOR INFO. SCI. & TECH. 132, 133 (2012) (explaining double-blind review and its elimination of some biases).

692. *See* Shobha Misra, *Randomized Double Blind Placebo Control Studies, the "Gold Standard" in Intervention Based Studies*, 33 INDIAN J. SEXUALLY TRANSMITTED DISEASES & AIDS 131 (2012) (describing use of randomization in clinical trials).

bias.[693] Applying double-blind review helps to keep partisanship from creeping into content moderation. If a moderator knows the identity of the politician, it may be difficult for the moderator to overcome any implicit bias against the politician the moderator has, no matter how well-intentioned. A 2018 study found that readers of news displayed greater partisanship and distrust of an article when they knew the source (versus not knowing the source).[694] For example, independents lowered their rating of trustworthiness of an article when a "liberal" source was disclosed; likewise, Democrats lowered their rating when it was disclosed the source was Fox News or Breitbart News.[695] While some observers may contend that such discounting of the source is justified for news consumption, in the context of content moderation on internet platforms the source of a post should not determine whether the content violates the community standards. If a moderator knows the identity of the source, then implicit bias might creep in. As Justice Ginsburg noted in a different context, blind auditions by orchestras to hire musicians helped to eliminate implicit bias against female musicians.[696]

A double-blind approach also helps to promote "good faith" moderation decisions under Section 230(c)(2) that are based on the material, and not the user's identity. And it helps to discourage politicians from attempting to curry favor from company executives or subject them to reprisals and intimidation if the politicians know that the executive can pull strings, so to speak, with the content moderators. This is not a hollow concern. Facebook executives have been criticized for their cozy relationships with Trump—and making decisions on content moderation based on the connection between Trump and Facebook executives.[697]

---

693. *See* Daniel Klerman & Greg Reilly, *Forum Selling*, 89 S. CAL. L. REV. 241, 254–55 (2016) (noting that the federal norm is "random assignment among judges within a district"); *cf.* RONALD A. FISHER, THE DESIGN OF EXPERIMENTS 19–21 (8th ed. 1971) (discussing randomization in experiments).

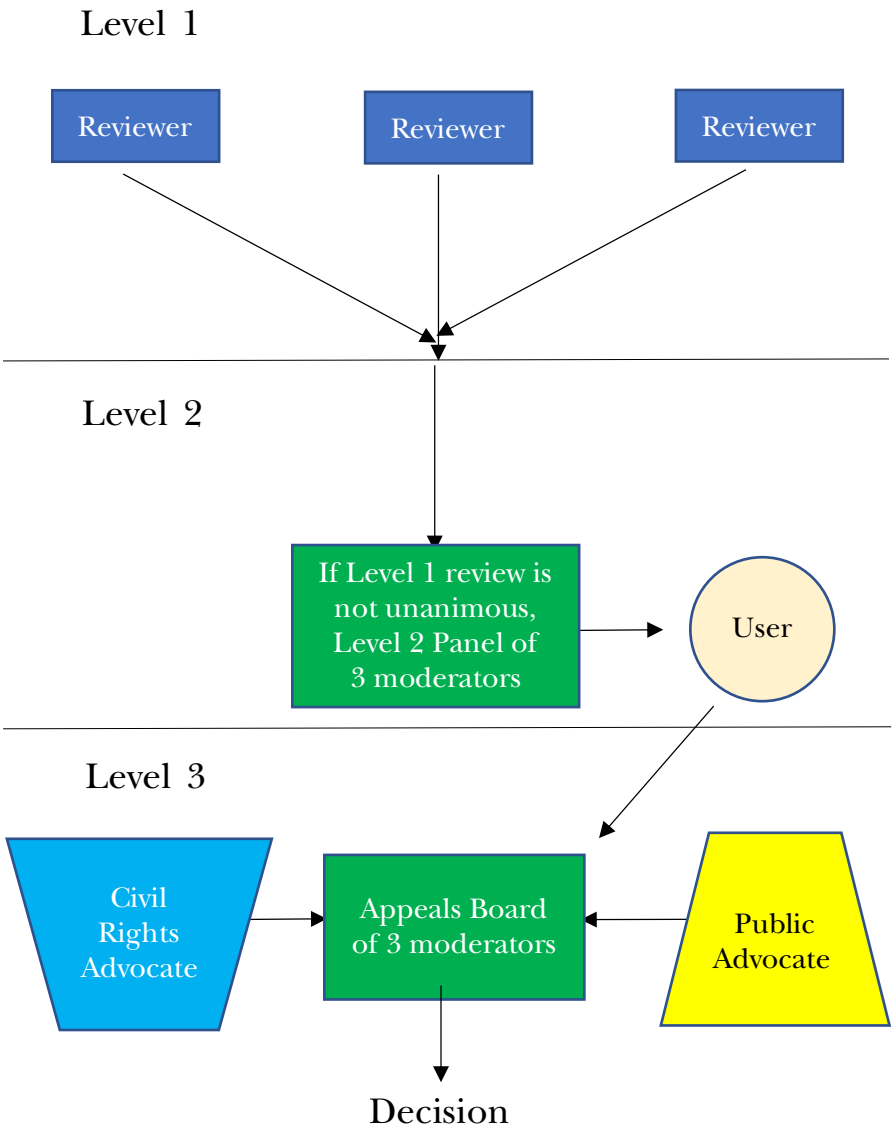
694. *See* KNIGHT FOUND., AN ONLINE EXPERIMENTAL PLATFORM TO ASSESS TRUST IN THE MEDIA 6 (2018), https://knightfoundation.org/wp-content/uploads/2020/02/KnightFoundation_NewsLens1_Client_Report_070918_ab.pdf [https://perma.cc/F6C8-PPMZ].

695. *Id.* at 8.

696. *See* Wal-Mart Stores, Inc. v. Dukes, 564 U.S. 338, 372–73, 373 n.6 (2011) (Ginsburg, J., concurring in part and dissenting in part).

697. *See, e.g.*, Dylan Byers & Ben Collins, *Trump Hosted Zuckerberg for Undisclosed Dinner at the White House in October*, NBC NEWS (Nov. 20, 2019, 11:32 PM), https://www.nbcnews.com/tech/tech-news/trump-hosted-zuckerberg-undisclosed-

Figure 1. Three Levels of Blind Review in the NCM Policy



dinner-white-house-october-n1087986 [https://perma.cc/F74V-V9BU]; Elizabeth Dwoskin et al., *Zuckerberg once Wanted to Sanction Trump. Then Facebook Wrote Rules that Accommodated Him.*, WASH. POST (June 28, 2020, 6:25 PM), https://www.washingtonpost.com/technology/2020/06/28/facebook-zuckerberg-trump-hate.

The use of three levels of review tracks the typical number of levels of courts available in the federal and state systems. Figure 1 above summarizes the three levels of review.

Level 1: Individual Moderators. The first level of review is conducted individually by three randomly assigned moderators, as explained in paragraphs 3 and 4 in the model policy above. Each determines if the (anonymized) content violates the community standard. There is no deliberation among the moderators in Level 1. If all three moderators reach the same decision, then a judgment is reached, and the three moderators agree on the penalty.

Level 2: Panel. However, if the moderators do not reach the same conclusion, then the case proceeds to Level 2. A new randomly assigned panel of three moderators is formed, as explained in paragraph 5. In Level 2, the review is different: the panel meets to discuss and then decides the case by a unanimous verdict. Although retaining the same moderators from the first level may save some resources, forming a new panel would provide a "second pair of eyes," with moderators who deliberate and discuss the case together before forming a final decision. Thus, the panel of moderators in the second level has no pre-commitment or anchoring bias to a decision the panel already formed. If the panel finds a violation, it decides what penalty to impose and notifies the user of the decision and the right to appeal it within a specified time.

Penalties. As Twitter and other platforms have shown, internet platforms have a variety of potential penalties or remedial actions to impose beyond simple removal of content (also known as takedown), which has the potential effect of altering political debate.[698] As evidenced in the controversy over Twitter's permanent suspension and Facebook's indefinite suspension of Trump's accounts following the January 6, 2021 attack on Congress, it behooves internet platforms to set forth, in advance, the precise policy or factors they use to determine when an account should be suspended. The model NCM policy above adopts an automatic rule of "3 strikes and you're out." If a user has committed three violations in the past three years, the user's account is automatically suspended. This categorical rule eliminates potential bias in suspension decisions. It also provides a clear rule—and potentially greater deterrence by putting users on

698. *See supra* notes 436, 446 and accompanying text.

notice.[699] Under the model NCM policy, the suspension is permanent, but the user can apply for reinstatement a year after the suspension if the user accepts "responsibility for the past violations" and "commit[s] to abide by the community standards." This reinstatement process is based on a recognition that a lifetime ban of a user from an internet platform is the ultimate penalty, carrying severe consequences that can affect both free expression and elections.[700]

Like most internet platforms have already done, Paragraph 7 recognizes a "public interest exception," by which the panel can exercise "discretion to decide to leave the content online after weighing the

699. *See generally* Willard K. Tom & Chul Pak, *Toward a Flexible Rule of Reason*, 68 ANTITRUST L.J. 391, 426 (2000) ("The clarity of the rules also makes it easier for the potential violator to conform its conduct to the rule and thereby improves deterrence independent of the severity of the penalties.").

700. The suspension decisions of Twitter and Facebook have been applauded and condemned. Without access to the internal exchanges between the companies and Trump for violations of their community standards leading up to the suspensions, I cannot fully evaluate the decisions. As discussed above, ideally companies should state the precise factors or rules they apply to determine if a suspension is warranted. Both Twitter and Facebook do so, employing an approach that appears to allow some discretion. *See About Suspended Accounts*, TWITTER, https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts [https://perma.cc/JYJ5-HA7U]; *Terms of Service*, FACEBOOK, https://www.facebook.com/terms [https://perma.cc/5TRS-YS7M]. Once discretion is allowed, it is more difficult to avoid the appearance of possible bias. Jack Dorsey, Twitter's CEO, expressed deep ambivalence and concerns with Twitter's suspension of Trump, but defended the decision as justified due to concerns about public safety. *See* Jack Dorsey (@jack), TWITTER (Jan. 13, 2021, 7:16 PM), https://twitter.com/jack/status/1349510769268850690. Facebook defended its indefinite suspension but referred the decision to the Oversight Board for review. *See* Nick Clegg, *Referring Former President Trump's Suspension from Facebook to the Oversight Board*, FACEBOOK (Jan. 21, 2021), https://about.fb.com/news/2021/01/referring-trump-suspension-to-oversight-board.

By contrast, the proposed NCM policy adopts an automatic rule of suspension if the user violates the policy three times in the preceding three years. This approach provides clearer warning to users when they face suspension of their accounts and eliminates potential bias in discretionary decisions. YouTube adopts a similar "3 strikes" approach, but it limits the window of strikes to just 90 days, meaning it is more forgiving than the proposed NCM policy. *See Community Guidelines Strike Basics*, YOUTUBE HELP, https://support.google.com/youtube/answer/2802032?hl=en [https://perma.cc/NSV3-KJCZ]. Under its policy, YouTube issued Trump a first strike on January 12, 2020, which led to the automatic 7-day suspension of Trump's YouTube channel during which he could not upload new videos. See Jennifer Elias, *Google Suspends Trump's YouTube Account, Barring Uploads and Comments*, CNBC (Jan. 13, 2021, 3:27 PM), https://www.cnbc.com/2021/01/12/google-suspends-trumps-youtube-account-disables-comments.html [https://perma.cc/68MG-DFJ7].

public's interest in viewing the content versus the potential harm in further dissemination." Other remedial actions short of removal include: adding a warning screen that a viewer must click to view the content, downgrading the search ranking of the content on the platform, and quarantining the content so it cannot be shared on the platform. Finally, in cases of repeated violations by the same user, the platform may consider suspension of the user account.

Level 3: Appeals Board. All verified political users have the right to appeal. If the user appeals the decision, the appeals board of three members reviews the appeal, deliberates together, and decides whether to affirm or reverse the decision. The user may challenge the finding of a violation of a community standard, such as by disputing that the content violates the standard or what the content means. The user may also raise a defense of selective enforcement as outlined below. During the appeal, a public advocate selected from a list of identified experts known for their professionalism, integrity, and expertise in relevant areas participates as amicus curiae to argue the interests of the public in the dispute. Congress recognized a similar public advocate in reforms to the secret Foreign Intelligence Surveillance Court in 2015.[701] Likewise, a civil rights advocate is also appointed to present the issues from a civil rights perspective in disputes involving hate speech, voter suppression, or other civil rights issues. The appeals board decides the dispute on the written submissions. In exceptional cases of profound public importance, the appeals board can hold an oral argument or public hearing. The appeals board then publishes a decision and the dispute is resolved.

5. Selective enforcement challenge by political candidate

This Article proposes a new type of challenge for selective enforcement of content moderation—to my knowledge, never before recognized by internet platforms. Under this defense, the user may challenge the violation based on a claim of selective enforcement due to partisanship or political affiliation. This type of challenge draws upon principles of the Fourteenth Amendment selective enforcement case law but sets forth a new type of defense and different requirements of proof. The model framework does not incorporate the same burden

701. *See* Cyrus Farivar, *America's Super-Secret Court Names Five Lawyers as Public Advocates*, ARS TECHNICA (Nov. 28, 2015, 7:00 AM), https://arstechnica.com/tech-policy/2015/11/americas-super-secret-court-names-five-lawyers-as-public-advocates [https://perma.cc/UJ2U-GUCK].

of proof that is required for a claim of selective prosecution under the Fourteenth Amendment, which requires the "a criminal defendant [to] demonstrate that the federal prosecutorial policy 'had a discriminatory effect *and* that it was motivated by a discriminatory purpose.'"[702] For practical reasons, requiring proof of intent in a corporate hearing for content moderation is unhelpful. It would be hard to square with double-blind review because the identities of the moderators might have to be disclosed to the challenger. Moreover, it could slow down the entire process given that the challenger may need discovery from the moderators who made the decision to understand their state of mind.

Instead of discriminatory intent, the following proof is required: the burden falls on the user to show that other verified political candidates or elected officials on the internet platform posted identical or substantially the same content, but were not subject to content moderation at the time the user's content was moderated. The public advocate may express a view on the validity of the user's assertion and may submit supporting or counter-evidence. And the appeals board may obtain from the company further analysis of treatment of similarly situated politicians for substantially the same content (if any). Based on the submissions, the appeals board will decide if the user has provided sufficient evidence of other similarly situated politicians who posted identical or substantially similar content, to prove selective partisan enforcement. If the appeals board finds selective enforcement, then it has two options: (1) allow the content to be reinstated without moderation or (2) moderate the content of both the appellant and the similarly situated politicians whose identical or substantially the same content also violated the community standards. Either way, the result is uniform.

6. Public advocate can appeal company's non-removal of content by political candidates

Under the proposed NCM framework, a public advocate will also be appointed when a panel in Level 1 or 2 decides that a political candidate has not violated the community standards. The public advocate will decide whether to appeal that decision to Level 3, and, if so, the political candidate will be afforded the opportunity to

---

702. United States v. Wallace, 389 F. Supp. 2d 799, 801 (E.D. Mich. 2005) (quoting Wayte v. United States, 470 U.S. 598, 608 (1985)).

participate in the appeal. This system is included because a company's decision not to remove a politician's arguably violating content may spark public scrutiny. Facebook's much ballyhooed Oversight Board is not currently authorized by Facebook to consider challenges to non-removal of content by Facebook, although Facebook suggested the Board's role could expand in the future.[703] This is a mistake, in my view. It results in an asymmetrical review system in which incentives are created for an internet platform to have a lax content moderation policy for political candidates (to avoid the expense of the elaborate procedures outlined above). More importantly, if community standards are to be meaningful and consistent, the same safeguards and procedures should apply to *non-removal* and removal decisions alike.

7. Comparison with Facebook's content moderation

The proposed model framework may sound elaborate and costly. However, it resembles some aspects of Facebook's own content moderation, as reported in Klonick's article.[704] Facebook uses three "tiers" of review: Tier 3 moderators from outsourced call centers overseas who make initial decisions, Tier 2 supervisors from the United States and also at the call centers who receive content that is "escalated" for further review or decide cases in which Tier 3 moderators disagreed, and Tier 1 moderators from the legal and policy department who have the final say if a case is escalated all the way up the review chain.[705]

Although the proposed model framework for moderation of political candidates also contains three levels of review, there are several important differences from Facebook's approach. First, every flagged post of a political candidate is guaranteed to receive at least three human moderators (Level 1). Second, none of the review is outsourced to call centers. All of the moderators are either employees or independent moderators chosen for their training, expertise, and integrity, particularly in the handling of politicians' content. Third, the model framework recognizes, in addition to a user's challenge to the determination of the violation, a defense of selective enforcement as a part of the appeals process. Fourth, the model framework requires, for

703. *See* Harris, *supra* note 681 ("As we continue to improve and expand the technology that makes appeals to the board possible, we want to also make it possible for people to refer cases where Facebook decided not to remove a piece of content.").

704. *See* Klonick, *supra* note 61, at 1640–42, 1647.

705. *Id.* at 1639–41.

every appeal, the appointment of an independent public advocate who is to represent the interests of the public before the board. If the content involves potential hate speech, voter suppression, or other civil rights issues, a civil rights advocate is also appointed to present the issues from a civil rights perspective. Fifth, the model framework allows appeals of the company's decisions not to remove content posted by a political candidate or public official that may violate a community standard.

B. Other Safeguards to Protect Against Partisan Content Moderation

Internet platforms should consider instituting other checks and balances to ensure nonpartisan content moderation of political candidates and ads.

1. Separation of powers: separation of content moderation from management and lobbying

Another important check on partisan content moderation that internet platforms should adopt is a principle of separation of powers.[706] There should be a figurative "wall" between the employees and executives tasked with content moderation and those responsible for management of the business, especially revenue-generation and lobbying. This is one of the demands sought by Rashad Robinson, President of Color of Change, which is a part of the coalition of civil rights groups leading the ad boycott against Facebook.[707] Importantly, the CEO of the company should not be involved in making content moderation decisions—or vetoing them. Otherwise, one person could undermine the checks and balances built into the multi-member and multi-level review for content moderation. In a court or administrative tribunal, it would be highly irregular for one person to be able to veto or override a decision from the proceedings below. True, a president as the executive has a veto power. But the veto power only applies to legislation. It does not extend to final judgments resolving disputes in our legal system. CEOs are highly

---

706. *See generally* Evelyn Douek, *Facebook's New 'Supreme Court' Could Revolutionize Online Speech*, LAWFARE (Nov. 19, 2018, 3:09 PM), https://www.lawfareblog.com/facebooks-new-supreme-court-could-revolutionize-online-speech [https://perma.cc/JH6R-CZX3] (characterizing Facebook's Oversight Board as a form of separation of powers).

707. *See* Sahil Patel, *Facebook Boycott Organizers Want a Civil Rights Expert in the Company's Executive Suite*, WALL ST. J. (July 2, 2020, 6:46 PM), https://www.wsj.com/articles/facebook-boycott-organizers-want-a-civil-rights-expert-in-the-companys-executive-suite-11593730015.

visible, and, if politicians know that CEOs are involved in or can veto content moderation decisions at their companies, the politicians can try to curry favor or even intimidate the CEOs. As noted above, Zuckerberg has been criticized for maintaining a cozy relationship with Trump.[708] Facebook board member and Trump-supporter Peter Thiel reportedly advised Zuckerberg against changing Facebook's policy to fact-check politicians.[709] And Zuckerberg reportedly decided to allow Trump's post about banning Muslims from immigrating to the United States, even though some employees contended that the posts violated Facebook's rule against hate speech.[710] The very appearance of a conflict of interest or partisan favoritism by the CEO in content moderation undermines the entire process. To some extent, Facebook may realize this problem given its willingness to establish the independent Oversight Board that will have the final say on a class of Facebook's content moderation decisions. Zuckerberg will not be able to veto the Board's decisions. In sum, internet platforms should recognize a principle of separation of powers in their community standards.

2. Transparency reports, independent audit, and expert advice

Internet platforms should include a specific section in their transparency reports devoted to content moderation of political candidates' content. Large internet platforms typically publish "transparency reports" online, detailing, with statistics, a host of practices, including their content moderation or community standards enforcement.[711] However, as of November 2020, these transparency reports typically do not identify their content moderation of political candidates or political ads, or, for that matter, election misinformation or voter suppression. Separate categories with statistics should be added for these variables and included in the transparency reports online.

Internet platforms should also consider having their content moderation practices subjected to a periodic, independent audit by a diverse group of relevant experts. Facebook enlisted an independent

---

708. *See* Byers & Collins, *supra* note 697.

709. *See* Glazer et al., *supra* note 35.

710. *See* Deepa Seetharaman, *Facebook Employees Pushed to Remove Trump's Posts as Hate Speech*, WALL ST. J. (Oct. 21, 2016, 7:43 PM), https://www.wsj.com/articles/facebook-employees-pushed-to-remove-trump-posts-as-hate-speech-1477075392?mod=e2tw.

711. *See, e.g.*, FACEBOOK, FACEBOOK TRANSPARENCY REPORT, https://transparency.facebook.com; RULE ENFORCEMENT, *supra* note 440.

"civil rights" audit led by Laura Murphy, a civil rights and civil liberties expert.[712] In July 2020, after a two-year review of Facebook, the auditors issued a highly critical report:

> While Facebook has built a robust mechanism to actively root out foreign actors running coordinated campaigns to interfere with America's democratic processes, Facebook has made policy and enforcement choices that leave our election exposed to interference by the President and others who seek to use misinformation to sow confusion and suppress voting.[713]

It is unclear whether Facebook will make changes in light of the audit,[714] but internet platforms should be open to doing so. Likewise, even short of conducting an extensive audit, the companies should consider enlisting the advice of a broad and diverse group of experts on issues related to content moderation and nonpartisanship, similar to TikTok's Content Advisory Council.

V. ADDRESSING CONCERNS WITH THE PROPOSED NCM FRAMEWORK

This Part addresses some of the major criticisms of the proposed NCM framework. While the concerns raised have some validity, they do not vitiate the overall goal of nonpartisan content moderation of political candidates or the general approach offered by the NCM framework, which is intended as a starting point for internet platforms to devise a more transparent and effective procedure of moderating the content of political candidates than currently exists.

A. Resources and Scalability

One concern is scale. As Twitter CEO Jack Dorsey remarked, content moderation "doesn't scale."[715] He was referring to the sheer volume of content on internet platforms. For example, by one estimate, Twitter has 350 thousand new tweets per minute, 500 million new tweets per

---

712. *See* Mike Isaac, *Facebook's Decisions Were 'Setbacks for Civil Rights,' Audit Finds*, N.Y. TIMES (July 8, 2020), https://www.nytimes.com/2020/07/08/technology/facebook-civil-rights-audit.html.

713. FACEBOOK'S CIVIL RIGHTS AUDIT—FINAL REPORT 10 (2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf [https://perma.cc/7VQ8-N9W8].

714. *See* Andrew Marino, *How far Will Facebook Go to Address Their Civil Rights Audit?*, VERGE (July 14, 2020, 3:50 PM), https://www.theverge.com/2020/7/14/21323988/vergecast-podcast-interview-rashad-robinson-color-of-change-facebook-ad-boycott.

715. *See* Aaron Beveridge, *Content Moderation "Doesn't Scale"—Jack Dorsey of Twitter*, YOUTUBE (Apr. 13, 2019), https://youtu.be/-w6oU33n_zU?t=71.

day, and 200 billion tweets per year.[716] Internet platforms already devote great resources and number of staff to content moderation. Facebook has 30,000 people devoted to content moderation (half are contract workers); YouTube, 10,000 people; and Twitter, 1,500 people.[717] Each internet platform must decide how best to operationalize additional safeguards to preserve nonpartisanship in the content moderation of political candidates and ads. One possibility is to take an incremental or pilot approach starting first with a smaller pool, such as (1) U.S. federal elections, (2) U.S. federal and state elections, or (3) U.S. federal, state, and local elections. As noted earlier, the number of politicians is 537 at the federal level, 18,749 at the state level, and 500,396 at the local level.[718] Even starting with one political office, such as the presidency, would be worthwhile. If successful, the pilot can be expanded.

B. Timeliness and Effectiveness Concerns

Another concern is that the intricate three levels of double-blind review in the NCM framework will be too cumbersome, time-consuming, and, ultimately, ineffective to avoid partisan review. To be sure, adding more due process to content moderation to ensure nonpartisanship will require time. Yet, with a pilot implementation of the NCM framework for a finite group of political candidates, even just starting with the presidency, a company's content moderation team can set forth a schedule to process the review in timely fashion. Moreover, the basic NCM framework is meant to be modular: it can be streamlined to two levels—decision and appeal—for expedited review. By comparison, the internet platforms have agreed to implement the EU's Code of Conduct against hate speech and "commit to reviewing the majority of these requests *in less than 24*

716. *See* Sayce, *supra* note 231.

717. *See* Elizabeth Dwoskin et al., *Content Moderators at YouTube, Facebook and Twitter See the Worst of the Web—and Suffer Silently*, WASH. POST (July 25, 2019, 1:00 AM), https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price; Brian Feldman, *Can 10,000 Moderators save YouTube?*, INTELLIGENCER (Dec. 5, 2017), https://nymag.com/intelligencer/2017/12/can-10-000-moderators-save-youtube.html.

718. *See supra* note 607 and accompanying text.

hours and to removing the content if necessary, while respecting the fundamental principle of freedom of speech."[719]

Another concern is that the NCM framework will create a disincentive for internet companies to moderate the content of politicians or political ads given the extensive levels of review proposed. Would it be easier just to leave content unmoderated? However, this incentive already exists under the current systems. At least under the NCM framework, users will be given a way to challenge non-removal decisions with the appointment of a public advocate to argue on their behalf. Thus, the NCM framework has a built-in mechanism to avoid an institutional disincentive against moderating content of political candidates. Moreover, given the concerns of election misinformation and voter suppression, internet platforms have reputational interests in not abdicating their responsibility of enforcing their community standards, especially related to content of political candidates and political ads.

A more troubling issue is that some, if not most, of the political candidates' content will contain information from which the moderators can determine the political party of the user who posted the content, as well as potentially even the user's identity. This type of content can be called "self-revealing" content. For example, even if the company removes the disclosure required by federal election law, a content moderator would know that an attack ad on Joe Biden likely originated from a Republican-supporting group. The moderator might even deduce the ad came from the Trump campaign, perhaps even mistakenly if the ad was created by a PAC not approved by Trump. The three levels of blind review in the NCM framework would not likely preserve the anonymity as to which political party the ad supports.

There is no easy way to avoid the problem presented by "self-revealing" content, even when it has been anonymized. Perhaps one additional safeguard to add to the NCM framework is the inclusion of employees who originate from other countries or moderators who do not follow U.S. national politics. (Moderators who do not know the candidates of state and local races would likely be easier to find.) Otherwise, the company must rely on the process (which includes a way for the candidate to raise a selective enforcement claim and the participation of a public advocate and potentially a civil rights advocate), as well as the

719. *The Code of Conduct on Countering Illegal Hate Speech Online*, EUROPEAN COMM'N (June 22, 2020) (emphasis added), https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135 [https://perma.cc/WL8P-6G2Q].

integrity of the moderators, to mitigate the potential for the biases of moderators to creep into their review of "self-revealing" content. But that is no different from the judicial system relying on the adversarial process and the jury and appeals systems to counteract potential bias tainting a verdict against a well-known defendant. Neither system is perfect.

C. Is Content Moderation Better Under the Status Quo than the NCM Proposal?

Finally, ardent supporters of Section 230 steadfastly defend Section 230 as being an important safeguard to the freedom of expression in that it allows user-generated content to be shared without exposing internet platforms to massive liability. Without Section 230, internet platforms would be more restrictive and start "censoring" user-generated content out of fear of being sued, such as for alleged defamation, a claim that is so easy to assert.[720] Or, even worse, platforms will stop publishing user-generated content altogether and switch to professional content.[721]

I share these concerns. But they do not address the main reason lawmakers seek reform of Section 230: the alleged political bias in how internet platforms moderate the content of candidates for public office and others. Strong proponents of Section 230 or the internet platforms may dismiss these allegations out of hand. The claims are either trumped up (where is the evidence of political bias?) or not problematic even if true (Fox News and CNN are just as politically biased). Let's stick with Section 230 and let the internet platforms carry on as usual. Congress is just wrong.

Even the proponents of Section 230 recognize, however, that Congress is poised to reform, if not outright repeal, Section 230. Indeed, Zuckerberg even asked Congress to do so in his testimony before the Senate Commerce Committee in October 2020:

> The debate about Section 230 shows that people of all political persuasions are unhappy with the status quo. People want to know that companies are taking responsibility for combatting harmful

720. *See* Derek E. Bambauer, *How Section 230 Reform Endangers Internet Free Speech*, BROOKINGS (July 1, 2020), https://www.brookings.edu/techstream/how-section-230-reform-endangers-internet-free-speech [https://perma.cc/SHM8-K86R].

721. *See* Eric Goldman, *An Interview on Why Section 230 Is on the 'Endangered Watch List,'* TECH. & MKTG. L. BLOG (Sept. 15, 2020), https://blog.ericgoldman.org/archives/2020/09/an-interview-on-why-section-230-is-on-the-endangered-watch-list.htm [https://perma.cc/XKQ9-NTA7].

> content—especially illegal activity—on their platforms. They want to know that when platforms remove content, they are doing so fairly and transparently. And they want to make sure that platforms are held accountable . . . . Changing it is a significant decision. However, I believe Congress should update the law to make sure it's working as intended.[722]

And, as noted above, President Biden is already on record in supporting a complete repeal of Section 230.[723]

Against this political landscape, the proposed NCM framework offers an alternative approach that internet platforms can undertake immediately to address the concerns of political bias—and to stave off Congress from shrinking Section 230's immunity beyond recognition.

CONCLUSION

The United States faces one of the most polarized moments in its history. This hyper-partisan climate has produced not only division and vitriol, but also a highly politicized attack on internet platforms' content moderation as showing an "anti-conservative bias" or, in the case of Facebook, showing favoritism to conservatives. Internet platforms, including Facebook and Twitter, face the threat of a dramatic reduction in the scope of immunity they can obtain under Section 230 of the Communications Decency Act under bills proposed by Republican lawmakers.[724] The companies categorically deny such a bias, but it is not entirely clear from their stated policies and community standards the precise "step-by-step" mechanisms that ensure nonpartisanship in their content moderation. To address this problem, this Article proposes a model nonpartisan content moderation (NCM) framework for internet platforms to adopt as a matter of best practices. The NCM framework provides greater transparency and institutionalized checks and balances and avoids messy entanglement of government enforcement of speech codes online. And it could help internet platforms avoid a complete overhaul or repeal of Section 230 immunity by Congress, while giving the public greater confidence that internet platforms are nonpartisan in their content moderation of political candidates.

---

722. Adi Robertson, *Mark Zuckerberg Just Told Congress to Upend the Internet*, VERGE (Oct. 29, 2020, 10:29 AM), https://www.theverge.com/2020/10/29/21537040/facebook-mark-zuckerberg-section-230-hearing-reform-pact-act-big-tech.

723. *See* Kelly, *supra* note 419.

724. *See supra* Section I.C.2 (summarizing proposed amendments to Section 230 to require political neutrality in content moderation).