Redrawing the Lines

An Assessment of the Impact of "Anti-Censorship" Legislation on Terrorist Content, Hate Speech/Harassment, and Mis/Disinformation

October 2022

Daveed Gartenstein-Ross, Madison Urban
& Cody Wilson

Author Biographies

Dr. Daveed Gartenstein-Ross is a scholar, author, practitioner, and entrepreneur who is the founder and CEO of Valens Global. He has been described by the director of the U.S. Department of Defense's Strategic Multilayer Assessment program as "the expert that the experts call to discuss the nettlesome challenges with terrorism and counterterrorism." In addition to leading Valens Global, Dr. Gartenstein-Ross serves as a senior advisor on asymmetric warfare at the Foundation for Defense of Democracies and an associate fellow at the International Centre for Counter-Terrorism – The Hague. He has previously held positions with the U.S. Department of Homeland Security, Google's tech incubator Jigsaw, and colleges and universities that include Carnegie Mellon University, Duke University, Georgetown University, and the University of Maryland.

Dr. Gartenstein-Ross led Valens Global's efforts to support the drafting of, and the threat assessment and crafting of priority actions for, the U.S. Department of Homeland Security's *Strategic Framework for Countering Terrorism and Targeted Violence*, which was released in September 2019. The *Strategic Framework* has subsequently guided the department's approach to confronting terrorism and other forms of violence. That document received widespread acclaim, with *The New York Times* editorializing that it "focuses unapologetically on right-wing terrorism, particularly white supremacist extremism," which constitutes "a shift that is both urgently needed and long overdue."

As a scholar, Dr. Gartenstein-Ross is the author or volume editor of over 30 books and monographs. He has testified before the Canadian House of Commons, European Parliament, U.S. Senate, and U.S. House of Representatives on relevant topics. He holds a Ph.D. in world politics from the Catholic University of America and a J.D. from the New York University School of Law.

Madison Urban is an analyst at Valens Global. In this role, she focuses on Valens Global's project examining domestic extremism. She has also contributed analysis relied on by top-tier law firms in high-stakes litigation and contributed to Valens Global's wargaming/simulations practice. Ms. Urban earned dual Bachelor's degrees in Public Policy and also Peace, War and Defense from the University of North Carolina-Chapel Hill.

Cody Wilson contributed to this study while working as an analyst at Valens Global, a position that he has now left. As a Valens analyst, Mr. Wilson worked on numerous Valens Global simulations, reports for high-stakes litigation, and projects for the U.S. and Canadian armed forces. He also contributed his technical

knowledge to the creation of a report examining information operations and to design a cybersecurity-focused tabletop exercise.

Mr. Wilson holds a master's degree in global studies and international relations, with a concentration in conflict resolution, from Northeastern University. He previously earned a bachelor's degree in political science with a concentration in international relations from the University of California, Los Angeles.

The authors have received support from Meta for their work on technology and national security.

Table of Contents

Introduction

On July 26, 2016, Abdel Malik Petitjean and Adel Kermiche stormed a church in Saint-Etienne-du-Rouvray, France. In an attack they credited to the Islamic State (ISIS) militant group, they took several hostages and slit the throat of an 86-year-old priest before being gunned down by police.[1] The two attackers had met in person for the first time only days earlier.

How did two people who barely knew one another come together to launch this grotesque plot? The answer is social media. Petitjean and Kermiche had been connected to one another by ISIS recruiter Rachid Kassim over the messaging app Telegram. As a company, Telegram was, at the time, frantically trying to deal with ISIS's use of its platform to propagandize, to recruit, and—as the Saint-Etienne-du-Rouvray attack shows—to plot. When the attack occurred, Telegram was taking down roughly 70 ISIS channels per day, culminating in almost 2,000 account takedowns per month as it tried to keep up with ISIS's abuse of its platform.[2] ISIS's other online innovations—including the manner in which it became world class at promoting its notorious "brand" over Twitter, and the organization's "virtual plotter" model that allowed it to facilitate attacks thousands of miles from its stronghold—are by now well known.

ISIS is one of the most prominent examples of the misuse of social media by malign actors, but other examples of this and similar kinds of abuse abound. Malign actors have used social media to spread terrorist content, engage in hate speech and harassment, and spread mis/dis-information. Social media companies have increasingly been called on to moderate content on their platforms to deal with these issues. **But tech companies' growing role in content moderation has not been without controversy**. The debate about content moderation and the role of Big Tech in society reached a crescendo following President Donald Trump's ban from Twitter and Facebook in January 2021. These bans were prompted by the companies' assessment that Trump's postings posed a risk of violence after his supporters stormed the U.S. Capitol on January 6. But even before these major platforms banned Trump, the anger of many political conservatives had been simmering due to perceptions that Big Tech's content moderation efforts were compromised by political bias. Trump's ban brought this anger to a boil.

The States of Texas and Florida, citing Trump's ban and other content moderation controversies, passed "anti-censorship" laws—Texas's House Bill (HB) 20 and Florida's Senate Bill (SB) 7072—seeking to constrain tech companies' ability to undertake content removal. Both laws were initially enjoined by federal courts, and subsequent legal wrangling now leaves the two laws in somewhat different places. In May 2022, the Eleventh Circuit Court of Appeals upheld the district court's preliminary injunction, while the Fifth Circuit Court of Appeals came to the opposite conclusion, lifting the preliminary injunction against the Texas law. The Supreme Court reversed the Fifth Circuit the same month, putting the preliminary injunction back in place. However, the Fifth Circuit then issued a ruling on the constitutional merits of HB 20 on September 16, 2022, declaring it constitutional and paving the way for a likely battle at the Supreme Court.[3]

---

[1] "French Church Attack: What We Know," *BBC News*, July 28, 2016, https://www.bbc.com/news/world-europe-36900761.

[2] ISIS Watch (@ISISWatch), December 26, 2016, Telegram post, https://t.me/ISISwatch/2.

[3] *NetChoice v. Paxton*, No. 21-51178 (5th Cir., September 16, 2022), p. 2, https://www.ca5.uscourts.gov/opinions/pub/21/21-51178-CV1.pdf.

Whatever merits the claims of tech companies' political bias may possess, if HB 20, SB 7072, or similar legislation had been in effect when the bulk of ISIS account removals occurred, they would have complicated efforts to counter the group's propaganda and recruiting. Nor is removal of terrorist content the only vital area related to malign actors' use of social media platforms that could be negatively impacted by legislation like HB 20 and SB 7072. Content moderation efforts pertaining to removing and containing hate speech, harassment, and mis/disinformation would all likely be negatively impacted.

There are legitimate concerns about possible political bias in companies' content moderation efforts, and companies have not always drawn the lines correctly in their decisions to restrict speech on their platforms. However, **HB 20 and SB 7072 go too far in imposing rigidity on platforms' content moderation**. This rigidity would be inflicted at a time when malign actors' tactics for exploiting online services are fluid, rapidly evolving, and increasingly dangerous.

The purpose of this report is to assess how legislative initiatives targeting social media content moderation could affect the ability of companies to control the prevalence of terrorist and extremist content, hate speech and harassment, and mis/disinformation on their platforms. The relative lack of regulation on companies has enabled them to respond in real-time to abuse of their services and adjust policies to counter malign actors' evolving tactics and techniques. Groups and individuals seeking to exploit social media platforms for malign ends are highly adaptable, have a keen sense of acceptable use policy boundaries, and aggressively test their limits. To keep up, social media companies rely on the flexibility and adaptability of content moderation policies. **Anti-censorship legislation such as SB 7072 and HB 20 threatens to disrupt this cat and mouse game, tipping the balance in favor of malign actors by weakening social media companies' responses**.

To understand the current push for anti-censorship legislation in Texas, Florida, and other states, it is important to know how we got here. The following section examines the ways social media companies have tried to counter misuses of their platforms. We then provide an overview of key laws relevant to the content moderation debate, including Section 230 of the Communications Decency Act, SB 7072, and HB 20. As part of this discussion, we examine the impacts of anti-censorship legislation across three lenses: 1) removal of terrorist and extremist content, 2) removal of hate speech and harassment, and 3) removal of polluted information.[4] We highlight how flexibility and innovation on tech companies' part have enabled them to respond to platform abuses in ways that prevented greater harms and likely loss of life. At the same time, these companies have not always drawn content moderation lines correctly, which has produced political controversies that now serve as the impetus for anti-censorship legislation. Overall, we conclude that HB 20 and SB 7072 would severely impair moderation of terrorist content, hate speech/harassment, and mis/disinformation in ways not intended or anticipated by the Texas and Florida legislatures.

[4] This report at times uses the term *polluted information* as a catch-all term for disinformation, misinformation, and malinformation. Polluted information muddies the information ecosystem. Each of the three other terms captured under the umbrella of polluted information have their own meanings. *Disinformation* is media that contains false or misleading information that is created and shared to intentionally harm a specific target. *Misinformation* contains false or misleading information but is not intentionally shared to cause harm; often the sharer is unaware that the media contains false or misleading information. *Malinformation* is media that contains true information, but sharing or spreading it would be harmful to a specific target. Personal photos, stories, and embarrassing information frequently constitute malinformation.

Abuse of Social Media Platforms & Resulting Content Moderation Policies

This section examines how tech companies have used their terms of service to remove dangerous and harmful content from social media platforms, and how they have blocked malign actors. Malevolent actors have frequently sought to circumvent takedown efforts and to build networks that disseminate noxious propaganda using social media. Corporate efforts to counter terrorist and extremist content, hate speech and harassment, and mis/disinformation have been largely reactive to these evolving threats. In seeking to protect their platforms, companies must weigh public safety imperatives against concerns about censorship, and as we noted, they have not always drawn the line perfectly. However, as this section demonstrates, the flexibility and adaptability of tech companies' content moderation policies have been critical to countering abuse of their platforms by actors with nefarious intent.

Terrorist and Extremist Content

On November 22, 2012, James Foley, an American freelance journalist covering the war in Syria, and his British colleague, John Cantlie, were captured by militants and held in captivity for eighteen months. On August 19, 2014, after he had been held and tortured, threatened, and starved for over a year and a half, James Foley was taken to Raqqa, where he was forced to kneel in front of a camera and recite a statement denouncing the United States before being beheaded by ISIS militants.[5] Amid the shock and disgust in the wake of Foley's beheading, efforts by Twitter and other platforms to stop the spread of the video became part of the national conversation.[6] Content moderation efforts targeting ISIS gained traction following the dissemination of footage of Foley's death on social media.

ISIS has been the most successful terrorist group to leverage social media for recruitment and mobilization to date. As one example of its proficiency, during the group's peak over 46,000 Twitter accounts were operated by the group's supporters from September to December 2014. With an average of 1,000 followers per account, ISIS was able to broadcast content to millions of people across the globe on Twitter alone.[7] In part due to the strength of ISIS's online communications, around 42,000 foreign fighters, hailing from over 120 countries, were drawn to fight with militant groups in Iraq and Syria.[8] This influx of foreign fighters eclipsed the number of fighters who joined the mujahedin in 1980s-era Afghanistan and the insurgency in mid-2000s Iraq.[9] The group also

[5] Their capture by militants is described in Rukmini Callimachi, "The Horror Before the Beheadings," *New York Times*, October 25, 2014, https://www.nytimes.com/2014/10/26/world/middleeast/horror-before-the-beheadings-what-isis-hostages-endured-in-syria.html. Full disclosure: one of this study's authors served as an expert witness in a civil lawsuit stemming from Mr. Foley's capture, torture, and execution. *Sotloff v. Syrian Arab Republic*, 525 F. Supp. 3d 121 (D.D.C., 2021).

[6] See David Weinberger, "Beheading Video Poses Challenge for Social Media," *CNN*, August 21, 2014, https://www.cnn.com/2014/08/21/opinion/weinberger-twitter-beheading/index.html; Bill Chappel, "Beheading Video Sets Off Debate Over How—or Whether—to Portray It," *NPR*, August 20, 2014.

[7] J.M. Berger & Jonathon Morgan, *The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter* (Washington, DC: The Brookings Institution, 2015), https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf.

[8] Joana Cook & Gina Vale, *From Daesh to 'Diaspora': Tracing the Women and Minors of the Islamic State* (London: International Centre for the Study of Radicalisation, 2018), https://icsr.info/wp-content/uploads/2018/07/Women-in-ISIS-report_20180719_web.pdf.

[9] Estimates of the number of foreign fighters in 1980s Afghanistan vary from 10,000 to 35,000, while estimates of the number of foreign fighters in mid-2000s Iraq vary from 4,000 to 5,000. See Peter Bergen, *The Osama bin Laden I Know: An Oral History of al-Qaeda's Leader* (Washington, DC: Free Press, 2006); Ahmed Rashid, *Taliban: Militant Islam, Oil*

tailored its social media recruitment toward specific skill sets, such as doctors, computer programmers, and media operatives.

In addition to bolstering ISIS's caliphate, social media was integral to the group's virtual plotter innovation, which effectively weaponized ISIS's propaganda efforts. In this model, operatives in ISIS's external operations division plot attacks online with supporters across the globe. The plotters provide logistical, tactical, and sometimes even emotional support to sympathizers seeking to carry out attacks. Prior to the advent of the virtual plotter model, this level of interaction between plotter and operative was reserved for face-to-face interactions.

One case that shows the advances made by the virtual plotter model is the Junead Khan plot, which was disrupted in July 2015. Khan had been on British authorities' radar since 2014, originally arousing suspicion for his desire to travel to the caliphate. In early 2015, he changed his mind and began focusing on carrying out a domestic attack, using his job as a deliveryman to scope out U.S. military bases. In July 2015, Khan and the virtual plotter with whom he was working, Junaid Hussain, discussed the logistics of various possible plans of attack. At one point, Hussain told Khan: "Most soldiers live in bases which are protected. I suppose on the road is the best idea."[10] Hussain later sent Khan a bomb making manual and told him to employ explosives against police who arrived on the scene of his attack.[11] While Khan may have previously been limited to passively reading about tactics and weapons, the virtual plotter model allowed Hussain to workshop Khan's attack plans with him and provide tailored tactical insights.

Mohammed Daleel's attack on a wine bar near a German music festival on July 24, 2016 further demonstrates the virtual plotter model's impact. Daleel was in direct contact with an ISIS virtual plotter from the planning stages of his attack until the moment of execution. In fact, absent Daleel's ongoing discussions with his handler, the attack may never have happened. As he scouted the target in the days before his attack, Daleel sent a picture to his handler, informing him that "this place will be crowded." Enthused, his handler replied: "Kill them all, so they'll be lying on the ground."[12]

As the day to attack arrived, however, Daleel was a bundle of nerves. The virtual plotter with whom he conversed helped him to overcome his doubts and redirect his attack:

> *Daleel*: "The party [concert] will be over soon, and there are checks at the entrance."
>
> *Virtual Plotter*: "Look for a suitable place and try to disappear into the crowd. Break through police cordons, run, and do it."

---

and Fundamentalism in Central Asia (New Haven, CT: Yale University Press, 2010); Thomas Hegghammer, "The Rise of Muslim Foreign Fighters," *International Security* 35:3 (Winter 2010/11), p. 61, https://www.belfercenter.org/sites/default/files/legacy/files/The_Rise_of_Muslim_Foreign_Fighters.pdf.

[10] "Luton Delivery Driver Guilty of Planning Terror Attack on U.S. Troops in Britain," *The Guardian* (London), April 1, 2016, https://www.theguardian.com/uk-news/2016/apr/01/luton-delivery-driver-junead-khan-guilty-planning-terrorattack-us-troops.

[11] "U.S. Airmen Terror Attack: Junead Khan Found Guilty," BBC (UK), April 1, 2016, https://www.bbc.com/news/uk35944661.

[12] Transcripts of Daleel's conversations with his handler can be found in "Auch der Attentäter von Ansbach Wurde vom IS per Chat Gestuert," *Süddeutsche Zeitung* (Germany), September 14, 2016, https://www.sueddeutsche.de/politik/terrordie-chats-der-attentaeter-von-wuerzburg-und-ansbach-mit-dem-is-1.3161419-2.

Daleel: "Pray for me. You do not know what is happening with me right now."

Virtual Plotter: "Forget the festival and go over to the restaurant [the wine bar]. Hey man, what is going on with you? Even if just two people were killed, I would do it. Trust in God and walk straight up to the restaurant."[13]

That is what Daleel did.

Heeding his handler's advice, Daleel detonated his bomb at a wine bar outside the concert, so he didn't have to face the security barriers he found daunting. The detonation killed Daleel and wounded 15 victims, four seriously.[14] While an operative who was on his own may have simply aborted the attack, Daleel's nerves were calmed in real-time by his handler.

Despite the proliferation of terrorist content on social media, tech companies were initially cautious to respond. One Twitter official was quoted anonymously in *Mother Jones* saying that "one man's terrorist is another man's freedom fighter" when asked if Twitter was going to get serious about pulling down ISIS content.[15] These words were spoken when James Foley and another American journalist had already been beheaded, when ISIS was engaged in a campaign of genocide against the Yazidi minority group, and when credible evidence and the group's own boasts showed that it was instituting sex slavery against captured women. Eventually, though, Twitter changed its course. The platform suspended over 1.2 million accounts for "terrorist content" between August 2015 and December 2017. Facebook removed 14.3 million "pieces of content related to ISIS, al-Qaeda, and their affiliates" in the first three quarters of 2018, while YouTube removed over 60,000 videos for violating its "policies against violent extremism" from September to December 2018.[16]

Such removals and suspensions had an impact on ISIS's reach. Twitter's suspensions proved detrimental to both the number of followers and amount of content associated with ISIS accounts. Even when accounts returned under a similar name, they struggled to gain the same amount of followers they enjoyed prior to their suspension.[17] The adaptability of social media companies' policies allowed them to mitigate the impact of ISIS recruitment on their platforms. As white supremacist extremists grew in prominence as terrorist threats, tech companies' terms of use and content moderation evolved as well, as evidenced by the response to the 2019 Christchurch mosque shootings.

---

[13] Ibid.

[14] Frederik Pleitgen, Tim Hume & Euan McKirdy, "Suicide Bomber in Germany Pledged Allegiance to ISIS Leader," CNN, July 26, 2016, https://www.cnn.com/2016/07/24/world/ansbach-germany-blast/index.html.

[15] Jenna McLaughlin, "Twitter Is Not at War With ISIS. Here's Why," *Mother Jones*, November 18, 2014, https://www.motherjones.com/politics/2014/11/twitter-isis-war-ban-speech/.

[16] Twitter Public Policy, "Expanding and Building #TwitterTransparency," April 5, 2018, https://blog.twitter.com/official/en_us/topics/company/2018/twitter-transparency-report-12.html; Monika Bickert, "Hard Questions: What Are We Doing to Stay Ahead of Terrorists?," Facebook, November 8, 2018, https://newsroom.fb.com/news/2018/11/staying-ahead-of-terrorists/; YouTube, "YouTube Community Guidelines Enforcement," (n.d.), https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?policy_removals=period:Y2018Q4&lu=policy_removals.

[17] J.M. Berger & Heather Perez, *The Islamic State's Diminishing Returns on Twitter: How Suspensions Are Limiting the Social Networks of English-Speaking ISIS Supporters* (Washington, DC: George Washington University, 2016).

On March 15, 2019, Brenton Tarrant massacred 51 people at two mosques in Christchurch, New Zealand, streaming the shooting on Facebook Live. The streaming of the attack "gave the footage the quality of a first-person 'shoot 'em up,'" as if the mass shooting was part of a video game.[18] Within the first 24 hours, Facebook removed 1.5 million videos of the shootings.[19] Work designed to stop proliferation of the video was described as "whack-a-mole" by one scholar, as the video was downloaded, cut, and reuploaded across a variety of platforms using techniques designed to avoid content recognition technology.[20]

In the months after the Christchurch attack, Facebook announced new policies, including a "ban on praise, support and representation of white nationalism and white separatism on Facebook and Instagram" and changes to its livestreaming policy that were designed "to innovate in the face of this threat" of abuse of streaming to disseminate hateful and violent content.[21] Platforms' ability to respond quickly to terrorist innovation and weaponization of social media has been a critical tool in limiting the proliferation of violent content.

Harassment and Hate Speech

In a study of online hate and harassment, the Anti-Defamation League found that 41% of Americans have experienced some type of online harassment and 27% have experienced severe online harassment, defined as "sexual harassment, stalking, physical threats, swatting, doxing and sustained harassment."[22] Despite social media companies' attempts to combat harassment and hate speech, these phenomena remain a significant problem. This section focuses the evolution of Facebook's Community Standards—including its policies on dangerous organizations and individuals, and on hate speech—and the ways Meta's standards governing content moderation and takedowns of individuals have thereby evolved.

Facebook began a broad-based, multi-year civil rights auditing process in 2018.[23] During this auditing process, the platform implemented the aforementioned ban on praise, support, and representation of white nationalism and separatism, which represented a shift from the previous and more limited ban against praise, support, and representation of white supremacy.[24] The company explained that its original more limited policy was based on distinguishing white nationalism from white supremacy because "we were thinking about broader concepts of nationalism and separatism — things like American pride and Basque separatism, which are an important part of people's identity." However, the company stated that ongoing conversations "with members of civil society

---

[18] Graham Macklin, "The Christchurch Attacks: Livestream Terror in the Viral Video Age," *CTC Sentinel* 12:6 (July 2019), https://ctc.usma.edu/christchurch-attacks-livestream-terror-viral-video-age/.

[19] Amy Gunia, "Facebook Tightens Live-Stream Rules in Response to the Christchurch Massacre," *Time*, May 15, 2019, https://time.com/5589478/facebook-livestream-rules-new-zealand-christchurch-attack/.

[20] Billy Perrigo, " 'A Game of Whack-a-Mole.' Why Facebook and Others Are Struggling to Delete Footage of the New Zealand Shooting," *Time*, March 15, 2019, https://time.com/5552367/new-zealand-shooting-video-facebook-youtube-twitter/.

[21] "Standing Against Hate," Facebook, March 27, 2019, https://about.fb.com/news/2019/03/standing-against-hate/; Guy Rosen, "Protecting Facebook Live from Abuse and Investing in Manipulated Media Research," Facebook, May 14, 2019, https://about.fb.com/news/2019/05/protecting-live-from-abuse/.

[22] *Online Hate and Harassment: The American Experience 2021* (New York: Anti-Defamation League, March 2021), p. 13, https://www.adl.org/media/16219/download.

[23] Laura W. Murphy & Megan Cacace, *Facebook's Civil Rights Audit – Final Report* (July 8, 2020), p. 5, https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf.

[24] "Standing Against Hate," Meta, March 27, 2019, https://about.fb.com/news/2019/03/standing-against-hate/.

and academics who are experts in race relations around the world have confirmed that white nationalism and white separatism cannot be meaningfully separated from white supremacy and organized hate groups."[25]

This expansion to include white nationalism and white separatism as categories of hateful expression prompted growth in the platform's moderation of broader categories of speech and expanded the number of people who could be removed from the platform. While Facebook originally allowed exceptions to its content moderation policy for humor-related content that contained white supremacist themes, the company refined its exception list to only permit "satirical" content after the civil rights audit noted problems with adequately defining humorous content.[26] The auditors noted that *satire* is a narrower category of speech with a more well-defined meaning, making it easier to enforce consistent content moderation standards, while *humor* is more subjective.

Facebook's decisions concerning combating harassment and hate speech sparked debates about what speech ought to be protected. Critics contend that hate speech is a subjective concept and there is a risk of moderators' bias influencing the implementation of such policies. Amid these policy changes and greater enforcement of Facebook's Dangerous Organizations and Individuals policy, Facebook banned Louis Farrakhan, Alex Jones, the conspiracy theory-oriented website *Infowars*, Milo Yiannopoulos, Paul Joseph Watson, Laura Loomer, and Paul Nehlen from its platforms in May 2019.[27]

While content moderation of openly violent actors, such as ISIS, had found greater approval, the suspension of these accounts was met with backlash and various allegations of bias, which are discussed in more detail subsequently. The inconsistent and somewhat ill-defined nature of what may trigger content moderation or removal has resulted in numerous folk theories (user-generated theories based on the user's perspective of how a platform works) about Big Tech's algorithms.[28] When content suddenly disappears, confusion sometimes abounds among users, the content creator, and sometimes even the platform. Content is sometimes automatically flagged for removal by an algorithm. If the content creator believes it was improperly removed, the process requires the content to be reviewed manually to adjudicate the appeal. During this process, content suitable for the platform is sometimes automatically flagged as inappropriate, leaving the content author and other users confused.

Another source of confusion stems from content that seems to disappear from public view without being expressly taken down, which is known in some online communities as *shadowbanning*. This folk theory argues that disappearing content is the result of content being algorithmically demoted, sometimes without clear reason. Shadowbanning, according to this theory, makes it harder for content to be discovered by other users, constituting a state that is less than an outright ban from the platform but that has a similar effect of making the user's content nearly invisible.[29] There are allegations across the political spectrum about bias and double standards in content moderation

---

[25] Ibid.
[26] Murphy & Cacace, *Facebook's Civil Rights Audit*, p. 44.
[27] Taylor Lorenz, "Instagram and Facebook Ban Far-Right Extremists," *The Atlantic*, May 2, 2019.
[28] See Brita Ytre-Arne & Hallvard Moe, "Folk Theories of Algorithms: Understanding Digital Irritation," *Media, Culture & Society* 43:807-824, https://doi.org/10.1177/0163443720972314.
[29] Caitlin Petre, Brooke Erin Duffy & Emily Hund, " 'Gaming the System': Platform Paternalism and the Politics of Algorithmic Visibility," *Social Media + Society* (October 2019), https://doi.org/10.1177/2056305119879995.

policies.[30] Hate speech can be difficult to define and despite posted policies, there is considerable opaqueness in the decision-making processes of tech companies, leading to confusion about what is and is not acceptable on platforms. This dynamic further reinforces allegations of bias.

Election-Related Influence Operations in 2016 and 2020

Russia conducted a major influence operation (IO) aimed at sowing political and social division in the United States in the run-up to the 2016 elections, much of which relied on social media. The Internet Research Agency (IRA) was primarily responsible for these activities. The IRA bought political advertisements, created social media posts designed to appear to be written by Americans, organized political events, and produced material disparaging Hillary Clinton.[31] To better understand political dynamics in the United States, IRA specialists had been studying American politics, influential groups, and trends, and the IRA began purchasing political advertisements in 2015, showing the operation's deliberate preparation.[32]

IRA-produced content reportedly reached hundreds of thousands if not millions of individuals, according to data from Facebook and Twitter that is cited in Special Counsel Robert Mueller's *Report on the Investigation into Russian Interference in the 2016 Presidential Election*. Examples of the operation's activities include running roughly 3,500 political ads on Facebook; scheduling competing rallies designed to get opposing sides in the same place at the same time, thus creating the possibility of conflict or violence; targeting ethnic minority groups with suggestions to vote for third-party candidate Jill Stein or stay home entirely; and creating a Twitter account that falsely claimed to be the official account of the Tennessee GOP.[33]

The efforts of Twitter, YouTube, Facebook, Google, and others to demote and remove "fake news" from their platforms following the 2016 elections spotlights one way that companies' algorithms can be used to suppress information. Yet as Hudson Institute scholar Harold Furchtgott-Roth points out, the idea of removing fake news can be problematic, as "one person's 'fake news' is another person's reality, and vice versa."[34] When tech companies determine what information is fake and what is not, critics worry that the suppression of information deemed by these companies to be untrue creates implicit limits on what constitutes acceptable speech. These platforms' pervasive role in public discourse has raised concerns about the effect that content moderation may have on free and open societies.

---

[30] For various writings on bias and double standards from both conservatives and liberal perspectives, see Kara Frederick, *Combating Big Tech's Totalitarianism: A Road Map* (Washington DC: Heritage Foundation, February 7, 2022); Ángel Díaz & Laura Hecht-Felella, *Double Standards in Social Media Content Moderation* (New York: Brennan Center, August 4, 2021); Christian Toto, "Big Tech's Double-Standard on Conservative Comics Is Getting Worse," *The Federalist*, March 4, 2021; Bharath Ganesh, "Content Moderation and Censorship: Can We Handle a Double Standard?," *OpenDemocracy*, May 29, 2019.

[31] Indictment, *United States v. Internet Research Agency*, 1:18-cr-00032-DLF (D.D.C., February 16, 2018), p. 4, https://www.justice.gov/file/1035477/download.

[32] Ibid., p. 12.

[33] The IRA's operation is laid out in exhaustive detail in both the Department of Justice's indictment of the organization as well as the Mueller report. The activities enumerated here constitute only a small sample of those conducted by the IRA.

[34] Harold Furchtgott-Roth, "Facebook, Google, and Twitter: Arbiters of Truth or Threats to Liberty?," *Forbes*, November 16, 2016, https://www.forbes.com/sites/haroldfurchtgottroth/2016/11/16/facebook-google-and-twitter-arbiters-ofthe-truth-or-threats-to-liberty/?sh=d3e13fa5ae38.

Due in part to the rise of foreign IOs and the opaqueness in content takedown policies, Facebook launched a fact-checking program in 2016 as part of an effort to "strike a balance between enabling people to have a voice and promoting an authentic environment."[35] The program alerted readers if the factual basis of an article was disputed, and the post distribution was subsequently reduced in users' feeds. Google News also began adding fact-checking labels in 2016, while Twitter has been slower to implement similar policies, and its fact-checking is limited to narrow categories.[36]

The suppression of an October 2020 *New York Post* article about Hunter Biden titled, "Smoking-gun email reveals how Hunter Biden introduced Ukrainian businessman to VP dad" is an oft-cited instance of bias against conservatives by social media companies. Indeed, it appears to be an example of legitimate speech that was suppressed. Twitter temporarily suspended the newspaper's account because of concerns that the article's exposé on the contents of Hunter Biden's computer hard drive violated its policy against distribution of hacked material. Facebook ensured that its algorithms limited the Hunter Biden laptop story's reach. This suppression was based not only on concerns about the distribution of hacked materials, but also due to assessments that the factual basis of the *New York Post* story was shaky, or perhaps even an IO by a foreign state. A number of conservative commentators slammed the action as a "political decision" made by Twitter to prevent negative coverage of Joe Biden in the weeks leading up to the 2020 election.[37] Illustrating that this suppression was questionable, various news companies, including *The Washington Post* and *The New York Times*, have now reversed their positions on the authenticity of the laptop.[38] This indicates that Twitter very likely drew the line incorrectly and suppressed valid information.

Again, though the line has demonstrably not always been drawn perfectly, these measures to flag and to redirect readers represent important evolutionary steps to counter polluted information. The relative effectiveness of Facebook's response to foreign influence operations on the platform can be discerned from the fact that its new policies had a concrete impact on the IRA's strategies. Analysis of the difference between Russia's influence campaigns in 2016 and 2020 highlights how the IRA adapted its posts' format and content to try to avoid detection by Facebook's algorithms. Posts during 2020 were more likely to contain text pasted verbatim from other sources to cut down on grammatical errors, used less hashtags, and contained more images in order to circumvent algorithms that primarily flag text.[39] Most significant from the perspective of these operations'

---

[35] "How Facebook's Third-Party Fact-Checking Program Works," Meta Journalism Project, June 1, 2021, https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works?locale=or_IN.

[36] "Google News Launches Fact Check Label," *BBC News*, October 14, 2016, https://www.bbc.com/news/technology-37657524; Yoel Roth & Ashita Achuthan, "Building Rules in Public: Our Approach to Synthetic & Manipulated Media," Twitter Blog, February 4, 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media; Kurt Wagner & Bloomberg, "Twitter Users Will Soon be Able to Report Misinformation for the First Time," *Fortune*, August 17, 2021, https://fortune.com/2021/08/17/twitter-users-combat-misinformation-tweets-social-media/.

[37] See, for example, Tim Murtaugh, "Election Panel's Giving Twitter a Pass on Hunter Biden Laptop Cover-up Is a Travesty," Heritage Foundation, September 15, 2021, https://www.heritage.org/election-integrity/commentary/election-panels-giving-twitter-pass-hunter-biden-laptop-cover (quoting the *New York Post*'s response to the FEC: "Does anyone think for a moment that Twitter would have censored a report about a laptop owned by Donald Trump Jr.?").

[38] See Craig Timberg, Matt Viser & Tom Hamburger, "Here's How the Post Analyzed Hunter Biden's Laptop," *Washington Post*, March 30, 2022; Katie Benner et al., "Hunter Biden Paid Tax Bill, But Broad Federal Investigation Continues," *New York Times*, March 16, 2022. See also discussion in Isaac Schorr, "The Washington Post Finally Gets Around to Confirming the Hunter Biden Laptop Story," *National Review*, March 30, 2022.

[39] Davey Alba, "How Russia's Troll Farm Is Changing Tactics Before the Fall Election," *The New York Times*, March 29, 2020, https://www.nytimes.com/2020/03/29/technology/russia-troll-farm-election.html.

changing efficacy, by 2020 the IRA generally sought smaller account followings in order to avoid suspicion.[40] However, foreign IOs continue to evolve to try to get around tech companies' monitoring. The ability of tech companies to evolve along with IO actors continues to be important.

The COVID-19 "Infodemic"

Throughout the COVID-19 pandemic, tech companies have faced both pressure to combat polluted information related to the virus, and also ongoing charges of bias in these efforts. The companies' efforts have suppressed legitimately dangerous misinformation, and in this way almost certainly saved lives. Yet the companies' policies have also, in other cases, resulted in the suppression of valid speech. Thus, tech companies' moderation actions during the pandemic simultaneously seem to provide support both for those who argue for greater moderation of polluted information as well as those who advocate for a more *laissez faire* approach.

On March 13, 2020, President Trump declared a national emergency due to the spread of the SARS-CoV-2 virus, which causes COVID-19.[41] Three days later, Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter, and YouTube released a joint industry statement committing themselves to combating misinformation on their platforms:

> We are working closely together on COVID-19 response efforts. We're helping millions of people stay connected while also jointly combating fraud and misinformation about the virus, elevating authoritative content on our platforms, and sharing critical updates in coordination with government healthcare agencies around the world. We invite other companies to join us as we work to keep our communities healthy and safe.[42]

In March 2020, UN Secretary General António Guterres commented on the prevalence of polluted information related to COVID-19, dubbing the phenomena an *infodemic*. He tweeted: "Our common enemy is #COVID19, but our enemy is also an 'infodemic' of misinformation. To overcome the #coronavirus, we need to urgently promote facts & science, hope & solidarity over despair & division."[43]

To help combat this "infodemic," tech companies stood up information centers, curated news feeds, and subtly redirected users from sources of polluted information to those information repositories. Meta (which was then known as Facebook) began to roll out pop-ups and information centers across Facebook, Instagram, and WhatsApp to direct and redirect people "to accurate information" about COVID-19. The company also engaged in content removal of "COVID-19 related misinformation that could contribute to imminent physical harm."[44] Meta also undertook efforts to limit the spread of posts with misinformation. Meta's Nick Clegg explained that "once a post is rated

---

[40] Ibid.

[41] Donald J. Trump, "Proclamation on Declaring a National Emergency Concerning the Novel Coronavirus Disease (COVID-19) Outbreak," March 13, 2020, https://trumpwhitehouse.archives.gov/presidential-actions/proclamation-declaring-national-emergency-concerning-novel-coronavirus-disease-covid-19-outbreak/.

[42] "Working With Industry Partners" Meta, March 16, 2020, https://about.fb.com/news/2020/12/coronavirus/.

[43] António Guterres (@antonioguterres), March 27, 2020, 11:55p.m. tweet, https://twitter.com/antonioguterres/status/1243748397019992065?s=20&t=zt5pPAN8vsyvGP1i7f7n-g.

[44] Nick Clegg, "Combating COVID-19 Misinformation Across Our Apps," Meta, March 25, 2020, https://about.fb.com/news/2020/03/combating-covid-19-misinformation/.

false by a fact-checker, we reduce its distribution so fewer people see it, and we show strong warning labels and notifications to people who still come across it, try to share it or already have."[45]

Twitter published its COVID-19 strategy on March 16, 2020. The strategy included an update to Twitter's definition of *harm*, and the company announced that it would act against content "that goes directly against guidance from authoritative sources of global and local public health information."[46] Violation of the guidelines could lead to the removal of tweets. Twitter also rapidly expanded its policy for placing notices on certain questionable tweets so that it could label misleading or disputed information related to COVID-19; when such labels were applied, the platform also provided links to information centers of curated content that were thought to be more accurate. The company also published the accompanying Figure 1, which delineates information or claims that would be prone to either labels or removal, based on their propensity for harm.[47]

These policies have almost certainly saved lives and safeguarded users' wellbeing by suppressing information about dangerous "cures." They have also at times suppressed legitimate speech. Turning first to examples of harm prevented by these policies, Twitter would remove tweets that included a "description of harmful treatments or preventative measures which are known to be ineffective or are being shared out of context to mislead people, such as 'drinking bleach and ingesting colloidal silver will cure COVID-19.'"[48] Miracle Mineral Solution (MMS), which is industrial bleach, was one such cure that was touted for COVID. The Grenon family manufactured, marketed, and sold MMS; they were later indicted after the "FDA received reports of people requiring hospitalizations, developing life-threatening conditions, and even dying after drinking MMS" despite FDA warnings against ingestion and previous court orders asking the family to cease operations.[49] Suppressing information about MMS and other questionable "cures" has almost certainly saved lives.

Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
Propensity for Harm		

Figure 1

However, at other times legitimate speech was suppressed by tech companies, including discussion of the idea that COVID-19 may have resulted from a lab leak in Wuhan, China. The claim that the COVID-19 virus might have originated in a Chinese lab in Wuhan was summarily dismissed as a conspiracy theory by tech companies and many prominent commentators early in the pandemic, with little apparent justification for this dismissal. For instance, in February 2021, Meta announced that it would remove false claims about COVID-19, including the claim that "COVID-19 is man-made or

[45] Ibid.

[46] Vijaya Gadde & Matt Derella, "An Update on Our Continuity Strategy During COVID-19," Twitter Blog, last updated April 1, 2020, https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.

[47] Yoel Roth & Nick Pickles, "Updating Our Approach to Misleading Information," Twitter Blog, May 11, 2020, https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

[48] "An Update on Our Content Moderation Work," Twitter, March 27, 2020, https://blog.twitter.com/en_us/topics/company/2020/covid-19.

[49] United States Attorney's Office for the Southern District of Florida, "Florida Family Indicted for Selling Toxic Bleach as Fake 'Miracle' Cure for Covid-19 and Other Serious Diseases, and for Violating Court Orders," April 23, 2021, https://www.justice.gov/usao-sdfl/pr/florida-family-indicted-selling-toxic-bleach-fake-miracle-cure-covid-19-and-other.

manufactured."[50] Eventually, the lab leak hypothesis became more widely accepted as a real possibility, including within the Biden administration, and social media companies' treatment of the theory largely mirrored this shift. Thus, three months after Meta initially announced its policy of suppressing the lab leak hypothesis, the company explicitly reversed course, stating that "in light of ongoing investigations into the origin of COVID-19 and in consultation with public health experts, we will no longer remove the claim that COVID-19 is man-made or manufactured from our apps."[51] While the origin of COVID-19 is still not known, and may never truly be known, the early suppression of a hypothesis that later received mainstream recognition as a valid possibility has been a source of legitimate concern.

Social media companies span multiple continents and their policies have the ability to impact the perception of billions, so discussions about the suppression of content—including harmful content on the one hand and legitimate speech on the other—are far from theoretical. Facebook's COVID-19 response team touts how it has connected 2 billion people across 189 countries with factual COVID-19 information, and how it has removed 15 million false claims.[52] Further, with at least half of Americans using social media for news at least sometimes in 2020, the impact of removing either potentially harmful information or legitimate speech should not be understated.[53]

The debate over how to hold social media companies accountable for both over-moderation and also under-moderation is complicated by the fact that **the very same mechanism—flexibility in defining the customer use policy—led to the initial suppression of the lab leak theory but also allowed tech platforms to be responsive to new information and reverse course as the hypothesis gained legitimacy**. Binding, inflexible guidelines might have stopped the suppression of legitimate speech initially, but also may have complicated Facebook's correction when new evidence came to light.

Political Controversies

Platforms' initial major push for moderation of terrorist content, primarily to counter pro-ISIS propaganda, rapidly gained bipartisan support in the United States. However, as content moderation expanded to cover more categories of purportedly harmful speech, these efforts became increasingly controversial. Though politicians of all stripes have been critical of Big Tech for one reason or another, content moderation has become a partisan issue. Democrats tend to favor aggressive moderation and want to hold social media companies *liable for content they choose not to moderate.* On the other hand, Republicans, particularly those close to Donald Trump, favor restrictions on content moderation and want to hold social media companies *liable for content they choose to moderate*. This report focuses on two bills passed by the Florida and Texas legislatures that mark the first state-level attempts to hold social media companies liable for the content that they moderate.

During the 2016 election campaign, both Hillary Clinton and Donald Trump called for tech companies to make the online environment inhospitable to jihadist militant groups. In a campaign

[50] Guy Rosen, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19," Meta, February 8, 2021, https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims.
[51] Ibid. (see update of May 26, 2021).
[52] "Responding to COVID-19," Meta, accessed February 22, 2022, https://about.facebook.com/actions/responding-to-covid-19/.
[53] Elisa Shearer & Amy Mitchell, "News Use Across Social Media Platforms in 2020," Pew Research, January 12, 2021, https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/.

speech at the University of Minnesota, Clinton advocated for tech companies to engage in a greater amount of content moderation:

> We have to stop jihadists from radicalizing new recruits in-person and through social media, chat rooms, and what's called the "Dark Web." To do that, we need stronger relationships between Washington, Silicon Valley, and all of our great tech companies and entrepreneurs. American innovation is a powerful force, and we have to put it to work defeating ISIS.… Companies should redouble their efforts to maintain and enforce their own service agreements and other necessary policies to police their networks, identifying extremist content and removing it.[54]

Trump also discussed the need to deny militants access to the internet, saying: "We have to go see Bill Gates and a lot of different people that really understand what's happening. We have to talk to them about, maybe in certain areas, closing that internet up in some way. Somebody will say, 'Oh freedom of speech, freedom of speech.' These are foolish people. We have a lot of foolish people."[55]

Russian interference in the 2016 election catalyzed a shift in the public conversation about social media companies' content moderation efforts. While there was bipartisan agreement in Congress about the need for content moderation to weed out bad actors trying to influence U.S. elections, at times this agreement gave way to disagreements about the impact of Russian interference, allegations of bias (principally made by conservatives), and declining trust in social media companies. Such factors increasingly made content moderation a hot-button issue.

Politicians on both sides of the aisle agreed that tech companies needed to identify and remove false ads and news propagated as part of Russian influence operations. In a 2017 Congressional hearing with the general counsels of Facebook, Twitter, and Google, Republican and Democratic senators called on the companies to do more. Senator Lankford (R-OK) commended the work of tech companies on terrorism, stating that they "have done a lot of work on terrorism, on Islamic extremism, on the advance of ISIS," but concluded that "we're asking for help on this area as well."[56] Senator Manchin (D-WV) highlighted the bipartisan nature of the request, saying that "this is not a Democrat or Republican issue. This is an American issue that we're concerned about, the security of our nation. We're getting hit from every way you possibly can imagine. And you all are the largest, one of the largest distributors of news. And there can be no doubt that it has to be authentic and true. You cannot allow what's going on against the United States of America."[57]

But even while many Republicans agreed with their Democratic colleagues that Russian efforts to influence the election were a national security threat, the pro-Trump wing of the Republican Party, including President Trump himself, largely felt that Russian interference was exaggerated for

---

[54] Hillary Clinton, "Remarks at the University of Minnesota in Minneapolis," December 15, 2015, https://www.presidency.ucsb.edu/documents/remarks-the-university-minnesota-minneapolis.

[55] David Goldman, "Donald Trump Wants to 'Close Up' the Internet," *CNN Business*, December 8, 2015, https://money.cnn.com/2015/12/08/technology/donald-trump-internet/.

[56] James Lankford, "Social Media Influence in the 2016 U.S. Election," Select Committee on Intelligence of the United States Senate, November 1, 2017, https://www.intelligence.senate.gov/hearings/open-hearing-social-media-influence-2016-us-elections#.

[57] Joseph Manchin, "Social Media Influence in the 2016 U.S. Election," Select Committee on Intelligence of the United States Senate, November 1, 2017, https://www.intelligence.senate.gov/hearings/open-hearing-social-media-influence-2016-us-elections#.

partisan reasons. In an interview with *Fox News*, president-elect Trump described the allegations that Russian interference helped him win the election as "'just another excuse" for why Clinton did not win, adding: "I don't believe it."[58]

Disagreements about the impact of Russian interference in the 2016 election also coincided with perceptions of platform bias against conservatives and declining overall trust in social media. As just one of the many examples of these allegations of bias, a May 2016 *Gizmodo* article gave voice to former Facebook employees who alleged that the company routinely suppressed conservative voices.[59] The percentage of Republicans who say they have a lot or some faith in the information that comes from social media platforms consequently dropped by 13 percentage points from 2016 to 2021, while Democrats exhibited only a 2-percent decline.[60]

Fears about bias in content moderation were further amplified when Facebook banned commentators Alex Jones, Milo Yiannopoulos, Paul Joseph Watson, Laura Loomer, and Paul Nehlen, as well as the website *Infowars* in May 2019. In response to the announcement, President Trump tweeted: "I am continuing to monitor the censorship of AMERICAN CITIZENS on social media platforms. This is the United States of America — and we have what's known as FREEDOM OF SPEECH! We are monitoring and watching, closely!!"[61] His next two posts amplified concerns about censorship of conservative voices, including a link to a *Breitbart* article titled "James Woods Banned from Twitter Amid Silicon Valley's Conservative Blacklisting Campaign."[62]

Despite such vehement objections, the overall commentary about these suspensions was far from negative. The parents of Noah Pozner, one of the victims of the Sandy Hook shooting, which Alex Jones has repeatedly called a hoax, had previously penned an open letter to Facebook asking for protection. They wrote: "Our families are in danger as a direct result of the hundreds of thousands of people who see and believe the lies and hate speech, which you have decided should be protected. What makes the entire situation all the more horrific is that we have had to wage an almost inconceivable battle with Facebook to provide us with the most basic of protections to remove the most offensive and incendiary content."[63] Senator Ron Wyden (D-OR) also praised Facebook's decision to remove these commentators, tweeting that he wrote Section 230 of the 1996 Communications Decency Act (discussed further later in this report) "so Facebook and other companies can take down bad actors without being saddled with frivolous lawsuits. Platforms need

---

[58] "Trump: Claims of Russian Interference in 2016 Race 'Ridiculous,' Dems Making Excuses," *Fox News*, December 11, 2016, https://www.foxnews.com/politics/trump-claims-of-russian-interference-in-2016-race-ridiculous-dems-making-excuses.

[59] Michael Nunez, "Former Facebook Workers: We Routinely Suppressed Conservative News," *Gizmodo*, May 9, 2016, https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006.

[60] Jeffery Gottfried & Jacob Liedke, "Partisan Divides in Media Trust Widen, Driven by a Decline Among Republicans," Pew Research, August 30, 2021, https://www.pewresearch.org/fact-tank/2021/08/30/partisan-divides-in-media-trust-widen-driven-by-a-decline-among-republicans/.

[61] Donald J. Trump (@realdonaldtrump), May 3, 2019 tweet, https://www.presidency.ucsb.edu/documents/tweets-may-3-2019.

[62] See Donald J. Trump (@realdonaldtrump), May 3, 2019, 7:23 p.m., Tweet, https://www.thetrumparchive.com/?dates=%5B%222019-05-03%22%2C%222019-05-04%22%5D; Donald J. Trump (@realdonaldtrump), May 3, 2019, 7:25 p.m., Tweet, https://www.thetrumparchive.com/?dates=%5B%222019-05-03%22%2C%222019-05-04%22%5D.

[63] Leonard Pozner & Veronique De La Rosa, "An Open Letter to Mark Zuckerberg: Our Child Died at Sandy Hook – Why Let Facebook Lies Hurt Us Even More?," *The Guardian* (London), July 25, 2018.

to be much more vigilant about weeding out hate. This is a start."[64] Others praised Facebook's move while cautioning against possible future abuses.[65]

The clamor of voices for, against, and taking a cautionary stance toward content moderation grew as the COVID-19 pandemic and 2020 election cycle further raised the stakes. As we discussed earlier, content moderation came to the forefront during the pandemic, as mis/disinformation related to COVID-19 posed the risk of causing serious injuries or claiming lives. Further, as the 2020 election approached, new policies had an impact on prominent figures. Controversy escalated when President Trump ran afoul of the new policies. On May 26, 2020, Twitter flagged one of Trump's tweets for the first time. The platform added a warning label with a link to an information center to one of the President's tweets that claimed mail-in ballots would be rigged.[66] The Trump campaign responded by accusing Twitter of political bias:

> We always knew that Silicon Valley would pull out all the stops to obstruct and interfere with President Trump getting his message through to voters. Partnering with the biased fake news media "fact checkers" is only a smoke screen Twitter is using to try to lend their obvious political tactics some false credibility. There are many reasons the Trump campaign pulled all our advertising from Twitter months ago, and their clear political bias is one of them.[67]

Allegations of bias and the fight over Twitter labels extended also to the platform's COVID-19 policies. In August 2020, for example, Twitter froze a Trump campaign account for violating its COVID misinformation policy. The Trump campaign described this move as a "display of Silicon Valley's flagrant bias against this President, where the rules are only enforced in one direction," as well as another occasion where tech companies were acting as "the arbiters of truth."[68]

Questions about how much social media companies should moderate public figures' accounts reached an apex after the Capitol riot on January 6, 2021. In the wake of the riot, Twitter announced Trump's permanent suspension from the platform for violating its glorification of violence policy, and due to concern that two particular Trump tweets "could inspire others to replicate violent acts," specifically "the criminal acts that took place at the U.S. Capitol on January 6, 2021."[69] One tweet read: "The 75,000,000 great American Patriots who voted for me, AMERICA FIRST, and MAKE AMERICA GREAT AGAIN, will have a GIANT VOICE long into the future. They will not be

---

[64] Ron Wyden (@RonWyden), May 2, 2019, 5:31 p.m., Tweet, https://twitter.com/RonWyden/status/1124063770802892803.

[65] For an example of one such warning, albeit in the context of the company's earlier ban of white nationalist and separatist content, see ACLU staff attorney Vera Eidelman's comments in Sasha Ingber, "Facebook Bans White Nationalism and Separatism Content From Its Platforms," NPR, March 27, 2019.

[66] Kate Conger & Davey Alba, "Twitter Refutes Inaccuracies in Trump's Tweets About Mail-In Voting," *The New York Times*, May 26, 2020, https://www.nytimes.com/2020/05/26/technology/twitter-trump-mail-in-ballots.html.

[67] Donald J. Trump, "Trump Campaign Statement on New Twitter Policy," The American Presidency Project, May 26, 2020, https://www.presidency.ucsb.edu/node/348527.

[68] Zack Budryk, "Twitter Bans Trump Campaign Until It Deletes Tweet with COVID-19 Misinformation," *The Hill*, August 5, 2020, https://thehill.com/policy/technology/510804-twitter-bans-trump-campaign-until-it-deletes-tweet-with-covid-19.

[69] Twitter Inc., "Permanent Suspension of @realDonaldTrump," January 8, 2021, https://blog.twitter.com/en_us/topics/company/2020/suspension.

disrespected or treated unfairly in any way, shape or form!!!"[70] The second tweet announced that Trump would not be attending Joe Biden's inauguration.[71]

Twitter's justification for Trump's permanent suspension explained the danger that the company discerned in these posts:

> President Trump's statement that he will not be attending the Inauguration is being received by a number of his supporters as further confirmation that the election was not legitimate and is seen as him disavowing his previous claim made via two Tweets (1, 2) by his Deputy Chief of Staff, Dan Scavino, that there would be an "orderly transition" on January 20th.
>
> The second Tweet may also serve as encouragement to those potentially considering violent acts that the Inauguration would be a "safe" target, as he will not be attending.
>
> The use of the words "American Patriots" to describe some of his supporters is also being interpreted as support for those committing violent acts at the US Capitol.
>
> The mention of his supporters having a "GIANT VOICE long into the future" and that "They will not be disrespected or treated unfairly in any way, shape or form!!!" is being interpreted as further indication that President Trump does not plan to facilitate an "orderly transition" and instead that he plans to continue to support, empower, and shield those who believe he won the election.
>
> Plans for future armed protests have already begun proliferating on and off-Twitter, including a proposed secondary attack on the US Capitol and state capitol buildings on January 17, 2021.[72]

Twitter thus determined that Trump's account should be banned because "the two Tweets above are likely to inspire others to replicate the violent acts that took place on January 6, 2021, and … there are multiple indicators that they are being received and understood as encouragement to do so."[73]

Facebook announced a 24-hour suspension of Trump's account on the night of January 6, 2021. That suspension became indefinite the next day.[74] The Trump suspension case was reviewed by Facebook's oversight board, an independent board that the company had established "to promote free expression by making principled, independent decisions regarding content on Facebook and

---

[70] Donald J. Trump (@realdonaldtrump), January 8, 2021, 2:46 pm, Tweet, https://www.presidency.ucsb.edu/documents/tweets-january-8-2021.

[71] Donald J. Trump (@realdonaldtrump), January 8, 2021, 3:44 pm, Tweet, https://www.presidency.ucsb.edu/documents/tweets-january-8-2021.

[72] Twitter Inc., "Permanent Suspension of @realDonaldTrump," January 8, 2021, https://blog.twitter.com/en_us/topics/company/2020/suspension.

[73] Ibid.

[74] Guy Rosen & Monika Bickert, "Our Response to the Violence in Washington," Meta, January 6, 2021, https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/.

Instagram and by issuing recommendations on the relevant Facebook company content policy."[75] The oversight board has the authority to make binding decisions about Facebook's content moderation. The board upheld the ban, and the company announced that Trump would remain suspended until at least January 2023.[76] Google joined in banning Trump's accounts later that week because of the "risk of incitement to violence," but has not provided a timeline for if or when he could be reinstated.[77]

Reactions to the Trump ban ranged from celebration to outrage. For many on the right, the ban was an example of a double standard by biased tech companies. Sen. Lindsey Graham (R-SC) highlighted a perceived double standard about the charge of incitement to violence: "Twitter may ban me for this but I willingly accept that fate: Your decision to permanently ban President Trump is a serious mistake. The Ayatollah can tweet, but Trump can't. Says a lot about the people who run Twitter."[78] Meanwhile, those on the left tended to see the ban only being imposed *after* the outbreak of violence as proof that tech companies needed to act more quickly. Sen. Mark Warner (D-VA) tweeted that Trump's ban was "an overdue step. But it's important to remember, this is much bigger than one person. It's about an entire ecosystem that allows misinformation and hate to spread and fester unchecked."[79]

Partisan divisions that had emerged around issues of deleting, suppressing, or flagging content are clear in the legislative record and in political commentary. While content moderation has been praised and explicitly requested by lawmakers on a bipartisan basis in certain cases, term-of-service enforcement and companies' policy changes have become contentious in recent years. One product of the increasing polarization surrounding this issue was the passage of SB7072 in Florida and HB20 in Texas, which shared the goal of limiting content moderation by social media companies.[80]

Legislation Affecting Content Moderation

This section seeks to untangle the increasingly complex legislative landscape affecting content moderation. At the federal level, Section 230 of the 1996 Communications Decency Act (CDA) affords social media platforms sweeping protections against being held liable for content moderation decisions. The following section examines key provisions of this law and demonstrates its impact on the evolution of content moderation. We also delve into the major controversies surrounding the law, which have contributed to a push for its repeal or alteration, as well as various state-level anti-

---

[75] The Oversight Board, homepage, accessed March 9, 2022, https://oversightboard.com/.

[76] Nick Clegg, "In Response to Oversight Board, Trump Suspended for Two Years; Will Only Be Reinstated if Conditions Permit," Meta, June 4, 2021, https://about.fb.com/news/2021/06/facebook-response-to-oversight-board-recommendations-trump/.

[77] Elizabeth Culliford & Paresh Dave, "YouTube Will Lift Ban on Trump Channel When Risk of Violence Decreases: CEO," Reuters, March 4, 2021, https://www.reuters.com/article/us-youtube-trump-suspension/youtube-will-lift-ban-on-trump-channel-when-risk-of-violence-decreases-ceo-idUSKBN2AW2LI.

[78] Lindsey Graham, @LindseyGrahamSC, Tweet, January 8, 2021, 8:15pm, https://twitter.com/LindseyGrahamSC/status/1347713459874627588?s=20&t=hYIgK-TBTXlHJJOt3Tdn-A.

[79] Mark Warner, @MarkWarner, Tweet, January 8, 2021, 6:47pm, https://twitter.com/MarkWarner/status/1347688913364783105?s=20&t=XWT7sQUfrm9JifnVL8D0SQ.

[80] Gov. Ron DeSantis, "Governor Ron DeSantis Signs Bill to Stop the Censorship of Floridians by Big Tech," May 24, 2021, https://www.flgov.com/2021/05/24/governor-ron-desantis-signs-bill-to-stop-the-censorship-of-floridians-by-big-tech/; Office of the Texas Governor, "Governor Abbott Signs Law Protecting Texans From Wrongful Social Media Censorship," September 9, 2021, https://gov.texas.gov/news/post/governor-abbott-signs-law-protecting-texans-from-wrongful-social-media-censorship.

censorship initiatives. The recently passed Florida and Texas laws, SB 7072 and HB 20, represent the leading edge of the anti-censorship movement, but there have been more than a dozen bills considered in state legislatures across the country since 2020 that target social media "censorship" and seek to place more restrictions on content moderation. Our discussion of the Florida and Texas bills in this section examines key provisions affecting content moderation in each law and highlights how malevolent actors could exploit provisions in each law to spread potentially dangerous and harmful content. We also summarize the current state of legal wrangling over both laws.

Section 230 of the Communications Decency Act

Section 230 of the Communications Decency Act, 47 U.S. Code § 230, affords internet-based content providers extensive federal liability protection for both allowing content on their sites and restricting it. The law's definition of internet-based content providers is broad, and seemingly encompasses blogs, social media companies, and even adult websites. Originally designed to enable the growth of the internet free from government interference and legal wrangling, Section 230 has become increasingly controversial on both sides of the aisle. Critics charge that the law is either too lax in permitting an "anything goes" virtual environment or is overly protective of providers' rights to moderate content. Section 230 has featured prominently in multiple court cases, and federal courts have repeatedly affirmed and often expanded the right to moderate content and Section 230's associated liability protections. Section 230 stands today as an obstacle to state-level anti-censorship legislation being enacted with full force, but with pressure building in Congress to increase regulation of tech companies, the future of Section 230 is uncertain.

Section 230 contains sub-section (c)(1), which states that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." This stipulation treats the host and uploader of content as separate entities, thus preventing a platform from being held legally liable for content produced or uploaded by its users. The law also contains important subsections protecting interactive computer services from liability for:

> (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).[81]

Taken together, these three provisions of the Communications Decency Act have enabled digital platforms to take as active or as hands-off an approach as they want in the moderation of content uploaded by users. This dynamic has produced bidirectional criticisms, as Section 230 has simultaneously been blamed for allowing too much content moderation and also not enough. Courts have thus far consistently interpreted Section 230 to both affirm the right of platforms to moderate content appearing on them and to protect platforms from liability for content they have chosen not

[81] 47 U.S.C. § 230.

to moderate.[82] Thus, individual platforms are provided with freedom to determine the approach they want to adopt.

Turning to the history of Section 230, the 1996 Communications Decency Act (CDA) was originally intended to restrict access to pornography and other graphic materials on the internet. However, large portions of the law pertaining to indecency were struck down by the Supreme Court as unconstitutionally vague in 1997.[83] Yet Section 230 remained in place, as the Court determined that the section could be severed from the rest of the CDA and remain in effect.

The original impetus for Section 230's addition to the CDA came after the landmark 1995 ruling in *Stratton Oakmont v. Prodigy Services*, in which the networked services provider Prodigy Services was found legally liable for content that it had failed to moderate.[84] The ruling treated Prodigy Services as though it were a publisher, thus putting the company on the hook for libelous content posted by its users. This ruling horrified other online service providers as well as lawmakers. Rep. Christopher Cox (R-CA), one of the co-authors of Section 230, said that he wanted to pass a law that would overturn the outcome of the *Stratton Oakmont* case. He noted that Prodigy's rival CompuServe had faced a similar lawsuit but was not found liable for libelous content on the platform because *it had never made an effort to moderate content*. Rep. Cox described the amalgamated outcome of the two cases as backward:

> The New York Supreme Court held that Prodigy, CompuServe's competitor, could be held liable in a $200 million defamation case because someone had posted on one of their bulletin boards, a financial bulletin board, some remarks that apparently were untrue about an investment bank, that the investment bank would go out of business and was run by crooks.
>
> Prodigy said, "No, no; just like CompuServe, we did not control or edit that information, nor could we, frankly. We have over 60,000 of these messages each day, we have over 2 million subscribers, and so you cannot proceed with this kind of a case against us."
>
> The court said, "No, no, no, no, you are different; you are different than CompuServe because you are a family-friendly network. You advertise yourself as such. You employ screening and blocking software that keeps obscenity off of your network. You have people who are hired to exercise an emergency delete function to keep that kind of material away from your subscribers. You don't permit nudity on

[82] See Cameron F. Kerry, "Section 230 Reform Deserves Careful and Focused Consideration," Brookings Institution, May 14, 2021, https://www.brookings.edu/blog/techtank/2021/05/14/section-230-reform-deserves-careful-and-focused-consideration/; *Jane Doe No. 1 v. Backpage.com, LLC*, 817 F.3d 12, 17 (1st Cir. 2016), p. 19, https://casetext.com/case/doe-v-backpagecom-llc-1.

[83] *Reno v. ACLU*, 521 U.S. 844, 875 (1997).

[84] Prodigy Services was once one of the "big three" online service providers, alongside Compuserve and America Online, that provided access to business, professional, and consumer information. Online service providers were integral to the early popularity of the internet in the 1990s by providing a range of information services, often with cutting-edge graphical interfaces that enabled easy navigation. Prodigy Services provided a "family-oriented" environment that offered news, weather, shopping, bulletin boards, games, and a variety of other services. For more information about the online environment and competition between information services in the mid-1990s, see Peter H. Lews, "The Compuserve Edge: Delicate Data Balance," *New York Times*, November 29, 1994, https://www.nytimes.com/1994/11/29/science/personal-computers-the-compuserve-edge-delicate-data-balance.html.

> your system. You have content guidelines. You, therefore, are going to face higher, stricter liability because you tried to exercise some control over offensive material."
>
> Mr. Chairman, that is backward. We want to encourage people like Prodigy, like CompuServe, like America Online, like the new Microsoft network, to do everything possible for us, the customer, to help us control, at the portals of our computer, at the front door of our house, what comes in and what our children see.[85]

Section 230 went into effect in February 1996. It aimed to remedy the problem that Rep. Cox articulated by providing protections for companies that made good faith efforts to moderate content.[86] According to the law's co-authors, Rep. Cox and Rep. Ron Wyden (D-OR), Section 230 was intended to encourage "Good Samaritan" self-regulation of content by platforms. As Rep. Cox explained on the floor of the House when introducing Section 230, the section was intended to

> do two basic things. First, it will protect computer Good Samaritans, online service providers, anyone who provides a front end to the internet, 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 internet, that we do not wish to have a Federal Computer Commission with an army of bureaucrats regulating the internet because, frankly, the internet has grown up to be what it is without that kind of help from the government. In this fashion we can encourage what is right now the most energetic technological revolution that any of us has ever witnessed. We can make it better. We can make sure that it operates more quickly to solve our problem of keeping pornography away from our kids, keeping offensive material away from our kids.[87]

Subsequent court rulings clarified the scope of the law's liability protection. Ultimately, courts have found that Section 230 protects a company's right to take as active or as hands-off of an approach as it wants to moderating user-generated content.[88] Christopher Cox, after leaving government, has argued that Section 230 became a "judge-made law" that is now inconsistent with the statutory intent, telling NPR that "he was shocked to learn how many Section 230 rulings have cited other rulings instead of the actual statute, [thus] stretching the law."[89]

---

[85] 141 Cong. Rec. H8469–70 (daily ed. Aug. 4, 1995) (statement of Rep. Christopher Cox).
[86] Danielle K. Citron & Benjamin Wittes, "The Problem Isn't Just Backpage: Revising Section 230 Immunity," *Georgetown Law Technology Review* 2:2 (2018), pp. 456-57, https://georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Citron-Wittes-453-73.pdf.
[87] 141 Cong. Rec. H8469–70 (daily ed. Aug. 4, 1995) (statement of Rep. Christopher Cox).
[88] See discussion in Cameron F. Kerry, "Section 230 Reform Deserves Careful and Focused Consideration," Brookings Institution, May 14, 2021, https://www.brookings.edu/blog/techtank/2021/05/14/section-230-reform-deserves-careful-and-focused-consideration/.
[89] Alina Selyukh, "Section 230: A Key Legal Shield for Facebook, Google Is About to Change," NPR, March 21, 2018, https://www.npr.org/sections/alltechconsidered/2018/03/21/591622450/section-230-a-key-legal-shield-for-facebook-google-is-about-to-change.

Section 230's broad liability protections have emboldened some companies to take an extreme hands-off approach to content moderation. One example is Backpage, an ad service that was repeatedly accused of facilitating sex trafficking.[90] On the other hand, other companies, like Twitter and Facebook, have shifted from a hands-off approach to more extensive moderation in the face of market and social pressures following platform misuse that promoted terrorist material, hate speech, and mis/disinformation, as previous sections of this report have discussed.

While Section 230 has been simultaneously critiqued by Republicans and Democrats, efforts to amend the provisions have usually been stymied from both sides of the aisle. The only successful effort thus far is the 2018 legislation known as FOSTA-SESTA, which combines the Senate's Stop Enabling Sex Traffickers Act (SESTA) and the House's Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA). FOSTA-SESTA's changes were rather limited, altering Section 230 to remove only the safe harbor protections for sex trafficking services.

Despite the widespread criticism of Section 230, it is worth noting that some commentators believe the provision should remain unchanged. TechDirt CEO Mike Masnick, for example, argues that Section 230 allows companies to respond effectively to market incentives.[91] A story praising Section 230 would proceed something like this. When the internet was first taking off in the 1990s, platforms were incentivized to take a hands-off approach to content moderation, and Section 230 allowed a vibrant ecosystem of expression and speech to thrive. As market pressures shifted, Section 230 also allowed tech companies to begin moderating their platforms, which was necessary to retain advertisers and to respond to misuse of the platforms. Section 230, Masnick argues, allows tech companies to be responsive to the stakeholders directly involved with the platform, including users, advertisers, and the public writ large, by choosing the level of moderation they deem appropriate.[92]

While Section 230 is the legal bedrock of the content moderation issue at present, its future is uncertain. As we have noted, there is significant political demand to reform or even repeal Section 230. As a general matter, if these laws were in contradiction to Section 230, it would be difficult for courts to determine whether federal law should trump state law or vice versa, and rulings would likely be inconsistent across jurisdictions. There are also clauses in the Florida and Texas laws that claim they are not intended to contradict federal law, but if that is the case, it isn't clear what their intended effect is (other than, perhaps, to send a message). We believe that an in-depth analysis of the Texas and Florida laws *as they are written* is helpful because state laws should not be crafted with the idea that contradictory federal laws will save us from their full effect.

An Overview of Florida Senate Bill (SB) 7072

Florida's SB 7072 was proposed by Governor Ron DeSantis in February 2021. DeSantis, a close ally of President Trump (who was, as this report explains, deplatformed from major social media platforms weeks earlier), unveiled the law as a way to "take aim" at tech companies that deplatform

[90] Ibid.
[91] Mike Masnick, "No, Internet Companies Do Not Get A 'Free Pass' Thanks To CDA 230," *TechDirt*, October 24, 2019, https://www.techdirt.com/articles/20191020/15092343224/no-internet-companies-do-not-get-free-pass-thanks-to-cda-230.shtml.
[92] Ibid.

political candidates.[93] The law passed by a wide margin in the Florida House of Representatives and by a narrower six-vote margin in the state senate.[94] Governor DeSantis signed the bill into law in May 2021, issuing a statement saying that the law's aim was to "hold Big Tech accountable by driving transparency and safeguarding Floridians' ability to access and participate in online platforms."[95] SB 7072 intended to achieve this goals by limiting how social media companies can remove users and user-generated content from their platforms. The law bars social media companies from deplatforming declared political candidates and requires these companies to publish standards for removing content and deplatforming individuals; apply those standards consistently; and notify users prior to any content removal, shadow banning, or deplatforming. The legislation also allows users to sue social media companies over certain practices deemed to be in violation of its key provisions.

At the heart of the law are two important sections relevant to content moderation by social media platforms that are also relevant to the law's potential impact on the three lenses this report focuses on: terrorist content, hate speech/harassment, and mis/disinformation. First is the creation of Section 106.072, titled "social media deplatforming of political candidates." This section states that

> a social media platform may not willfully deplatform a candidate for office who is known by the social media platform to be a candidate, beginning on the date of qualification and ending on the date of the election or the date the candidate ceases to be a candidate. A social media platform must provide each user a method by which the user may be identified as a qualified candidate and which provides sufficient information to allow the social media platform to confirm the user's qualification by reviewing the website of the Division of Elections or the website of the local supervisor of elections.

The section goes on to state that "this section may only be enforced to the extent not inconsistent with federal law and 47 U.S.C. s. 230(e)(3), and notwithstanding any other provision of state law." This line references a provision of Section 230 that allows states to enforce laws relating to content moderation only so long as those state laws are consistent with Section 230. Despite this provision, observers remain uncertain about how Section 230 and the Florida law would intersect in practice if the latter is not blocked by the courts.[96]

Perhaps the most critical anti-censorship component of SB 7072 is Section 501.2041, which states that failure to comply with this portion of the law could constitute an "unfair or deceptive act."[97]

---

[93] "Gov. DeSantis Proposes Law That Would Fine Big Tech Companies that 'Deplatform' Political Candidates," *WFLA 8*, February 2, 2021, https://www.wfla.com/news/florida/gov-desantis-holds-press-conference-from-florida-state-capitol/.

[94] "FL S7072 | 2021 | Regular Session," LegiScan, May 25, 2021, https://legiscan.com/FL/bill/S7072/2021.

[95] Governor Ron DeSantis, "Governor Ron DeSantis Signs Bill to Stop the Censorship of Floridians by Big Tech," May 24, 2021, https://www.flgov.com/2021/05/24/governor-ron-desantis-signs-bill-to-stop-the-censorship-of-floridians-by-big-tech/.

[96] See, for example, Devin Coldewey, "Florida's Ban on Bans Will Test First Amend Rights of Social Media Companies," *TechCrunch,* May 24, 2021, https://techcrunch.com/2021/05/24/floridas-ban-on-bans-will-test-first-amendment-rights-of-social-media-companies/; Jerry Lambe, "Florida Social Media Censorship Law Touted by Gov. DeSantis Is a 'Frontal Assault on the First Amendment': Lawsuit," *Law & Crime,* May 27, 2021, https://lawandcrime.com/lawsuit/florida-social-media-censorship-law-touted-by-gov-desantis-is-a-frontal-assault-on-the-first-amendment-lawsuit/.

[97] The determination that a company has engaged in an "unfair or deceptive act" carries several possible penalties beyond the civil liability discussed in the following pages. As explained in § 287.137 of the Florida law, companies found

The section goes on to stipulate that social media companies must publish the standards and definitions the platform "uses or has used for determining how to censor, deplatform, and shadow ban." The law continues that companies "must apply censorship, deplatforming, and shadow banning standards in a consistent manner among its users on the platform," must inform users about any changes to its rules ahead of time, and cannot change their rules "more than once every 30 days." Social media companies are also prevented from removing content, shadowbanning, or deplatforming without providing a user with prior notice, unless the material can be classified as obscene under Florida law, in which case no prior notification is needed.

In addition, § 501.2041 requires that users have access to metrics of who their posts were shown to, and requires platforms to categorize curation algorithms.[98] The law also allows users to opt out of those algorithms, requires platforms to provide users with an annual report on curation algorithm usage, and prohibits curation of political candidate posts unless that candidate pays for post promotion. Additionally, platforms must allow users access to their data if they have been deplatformed, prohibits the censorship or deplatforming of journalistic enterprises, and lays out extensive requirements for notifying users of why their posts were taken down.

A core concern for many observers is sub-section (6), which allows users to bring legal action against social media companies for inconsistent "censorship, deplatforming, and shadow banning standards" or for "censoring, shadow banning, or deplatforming" a user without prior notice. This part of the law gives users residing in Florida the ability to allege in a lawsuit that a social media company unfairly removed or silenced them or their content. A victorious plaintiff would be eligible for several remedies, including up to $100,000 in statutory damages per proven claim, punitive damages (if aggravating factors are present), attorney's fees, and injunctive relief. This section of the law concludes by again highlighting that it should not be construed to be inconsistent with or invalidate relevant federal laws.

Calls for or anticipation of legal challenges to SB 7072 emerged quickly, even before DeSantis signed the bill.[99] On May 27, 2021, two tech industry trade associations, NetChoice and the Computer & Communications Industry Association (CCIA), filed a lawsuit in the U.S. District Court for the Northern District of Florida arguing that the Florida bill infringed on various free speech

---

to be engaged in unfair or deceptive practices could be labeled antitrust violators, which could cause the company to be blacklisted from doing business in the state of Florida. For additional information about the antitrust framework, see S.B. 7072, Sess. of 2021 (Fla. 2021), https://www.flsenate.gov/Session/Bill/2021/7072/BillText/er/HTML.

[98] *Curation algorithms* are artificial intelligence-driven processes intended to provide the user with a better experience. Curation algorithms are frequently designed to learn what a user likes and enjoys, and to make more of that content available to the user. Implementations of a curation algorithm can also rank certain material that is deemed "high quality" by a series of internal metrics and offer that content above other content. Google's search engine, for example, does this by ranking pages, or PageRank, based on several factors, including a website's importance based on how often users interact with that site, and how well it fits the search criteria. More recently, social media companies curated news related to COVID-19, favoring material from authoritative sources like the Centers for Disease Control, the World Health Organization, or local public health departments.

[99] For examples, see Leandra Bernstein, "Florida State Officials Draft Legislation on Big Tech Censorship," *ABC6*, February 3, 2021, https://abc6onyourside.com/news/nation-world/florida-state-officials-draft-legislation-on-big-tech-censorship; Corbin Barthold & Berin Szoka, "No, Florida Can't Regulate Online Speech," *Lawfare*, March 12, 2021, https://www.lawfareblog.com/no-florida-cant-regulate-online-speech; Devin Coldewey, "Florida's Ban on Bans Will Test First Amend Rights of Social Media Companies," *TechCrunch*, May 24, 2021; Gilad Edelman, "Florida's New Social Media Law Will be Laughed Out of Court," *Wired*, May 24, 2021, https://www.wired.com/story/florida-new-social-media-law-laughed-out-of-court/.

protections and was also politically motivated.[100] The court subsequently issued a preliminary injunction on June 30, 2021, which the Eleventh Court upheld on May 23, 2022. The State of Florida then asked the U.S. Supreme Court to review the matter, filing a petition for writ of certiorari.[101]

Considerations Based on the Law's Text

While SB 7072 is intended to "hold big tech accountable," the law is plagued by several loopholes that malign actors could exploit. We now discuss some of the ways that SB 7072 could be used by such individuals and groups.

As an initial matter, it is worth noting that SB 7072 contains a provision stating that it shall not disrupt content moderation of material that violates state or federal laws. Such a provision is insufficient to prevent the law from interfering with social media companies' efforts to engage in content moderation to deal with terrorist content, hate speech/harassment, and mis/disinformation on their platforms. Categories of outright illicit material, *even as it relates to such illicit ends as terrorism*, are vastly outnumbered by licit material. That is to say, most harmful content related to terrorism, hate speech, harassment, or mis/disinformation is not illegal. Hate speech, for example, is perfectly legal, as is recruiting for violent extremist groups so long as they are not designated as terrorist groups under federal law. For example, recruiting for the violent white supremacist groups Atomwaffen Division or The Base is technically legal, since these groups are not designated terrorist organizations. This renders the statutory carveout enabling the removal of illegal material much less effective than one might hope.

Further, SB 7072 forces social media companies to publish their entire playbook for how they prevent misuse of their platforms, which malicious actors can then access, study, and exploit. An actor can figure out how to just barely stay within the lines of a company's policies, or the actor can come up with novel approaches that are not covered by a content moderation standard that Florida's law renders inflexible. For example, imagine a neo-Nazi in a small town who wants to target the local Jewish population, so he decides to fly a drone just outside the property of each Jewish resident and stream video footage from the drones to Facebook. There is unlikely to be a policy that explicitly covers this situation since the victim's property rights are not being violated, nor is their personal information necessarily being leaked online (a recognized form of harassment known as *doxxing*). In response, Meta would have to create a new rule banning the activity, as the behavior is clearly intended as harassment, intimidation, and (at the least) an implied threat. Yet the process of re-promulgating rules will likely take time since Meta is a large company with many bureaucratic hurdles to go through for new policies to be approved. Further, Meta would have to notify all users of this policy change prior to enforcing it, and by the terms of the Florida law, the company *would be unable to promulgate the new rule if it had already effected a rules change in the previous 30 days*. Thus, the Florida law turns what would have been a relatively easy fix under the old system, which

---

[100] "Big Tech Associations Sue Florida Over New Social Media Censorship Law," *WFLA 8*, May 27, 2021, https://www.wfla.com/news/politics/big-tech-associations-sue-florida-over-new-social-media-censorship-law/.

[101] *Netchoice LLC v. Moody*, 4:21cv220-RH-MAF (N.D. Fla., Jun. 30, 2021), pp. 30-31, https://storage.courtlistener.com/recap/gov.uscourts.flnd.371253/gov.uscourts.flnd.371253.113.0.pdf; *NetChoice v. Moody*, No. 21-12355 (11th Cir., May 23, 2022), p. 4, https://media.ca11.uscourts.gov/opinions/pub/files/202112355.pdf; See *NetChoice v. Paxton*, 596 U.S. ___ (2022); Petition for Writ of Certiorari, Moody v. NetChoice, No. (11th Cir., September 21, 2022), https://netchoice.org/wp-content/uploads/2022/09/Moody-FL-v.-NetChoice-Petition-for-Certiorari-filed-Sept-21-2022.pdf.

provided Meta with flexibility, to an incredibly cumbersome process. In the meantime, very real consequences could befall the victims.[102] Further, if the neo-Nazi streamer were to fashion his harassment as a journalistic enterprise, the Florida law would prevent it from being pulled down at all due to the provision that prohibits the censorship or deplatforming of journalistic enterprises.

An extremist group styling its work as journalism to avoid takedowns of its content is far from a theoretical problem. Amaq News Agency, one of ISIS's premier propaganda outlets, has long styled itself as reporting objectively on ISIS's activities rather than cheerleading for them. Other terrorist groups have had or experimented with their own journalistic outlets, including al-Qaeda, FARC (prior to its 2021 removal from the U.S.'s terrorist list), Hamas, and Hizballah. Indeed, under SB 7072 a "journalistic enterprise" simply means any entity doing business in Florida that posts 100,000 words online to an audience of 100,000 of users per month, or that posts a mere 100 hours of audio or video for 100 million viewers in a year.[103] At its peak, Amaq would have easily been able to produce 100,000 words covering ISIS's activity per month; given the number of ISIS supporters, as well as professional ISIS watchers (e.g., media, academics, experts, security professionals), a readership of 100,000 per month would have been trivial for it to maintain.

While media outlets that are clearly a part of designated terrorist organizations could be deplatformed under the framework established by SB 7072, terrorist groups have experience setting up front groups specifically designed to obscure their connection to proscribed groups. It can be difficult even for nation-states to discern when a front group is part of a terrorist organization, and it would be even harder—and more resource-intensive—for a tech company to pursue such an investigation to the point where its proof would satisfy a court of law. The path of least resistance for tech companies under SB 7072 would be to allow faux journalist enterprises—all of which are potential plaintiffs—to propagandize for terrorist groups under the pretext that they are not connected to the militant organizations. This is particularly true in the context of white supremacist groups, as only a single group that is considered a part of that milieu—the Russian Imperial Movement—has been placed on a terrorist list by the U.S. government. For example, the neo-Nazi website *Stormfront* brought in over 300,000 users in January 2022 alone, and is already headquartered in West Palm Beach, Florida, so the extra step of establishing itself as an entity doing business in Florida would be unnecessary: *Stormfront* could already qualify as a journalistic enterprise under SB 7072.[104] **The threshold for being considered a journalistic enterprise under SB 7072 would be relatively easy for almost any terrorist, violent extremist, or hate group to reach if it created a "news" website dedicated to covering its own activities.** This is a massive loophole that would allow SB 7072 to be exploited to promote terrorist content and other kinds of content that the bill's sponsors no doubt did not intend.

Nor is this the only issue with SB 7072. The protection that the law affords to political candidates could also be exploited by malign actors. Under existing Florida election laws, all a person has to do to qualify as a political candidate in Florida is apply to be a write-in candidate.[105] Thus, all extremists of any stripe would need to do to gain protection from "censorship" under SB 7072 is register as a

[102] In this example, another neo-Nazi might decide to use the video to locate and follow victims, or to attack them on stream with the hope of being immortalized in neo-Nazi online culture.
[103] The law also includes cable channels and businesses that have an FCC broadcast license as "journalistic enterprises." S.B. 7072, Sess. of 2021 (Fla. 2021), https://www.flsenate.gov/Session/Bill/2021/7072/BillText/er/HTML.
[104] "Stormfront.org," *Similarweb*, January 2022, https://www.similarweb.com/website/stormfront.org/#overview.
[105] This is simply the easiest way; there are also several other ways to qualify as a political candidate.

candidate seeking election via write-in with the Florida Division of Elections or a local election supervisor. The person could then even purchase advertising that the social media companies would be forced to host! The American Nazi Party, as just one example, could gain protection from content moderation for its political candidates. Similar legal loopholes have been abused by white supremacists previously. In 1985, white supremacist William Pierce, author of the notorious book *The Turner Diaries*, founded the religion of Cosmotheism and an accompanying Cosmotheist Community Church as a way of gaining tax exempt status for his white supremacist activities.[106] Similarly, a person could exploit the loopholes in SB 7072 to register as a political candidate and then run on a pro-Nazi or pro-ISIS platform—and companies like Meta, Twitter, and YouTube would be legally barred from removing their posts.

SB 7072 also gives the Florida Attorney General and Floridians the ability to bring legal action in court for violation of the law. This cause of action is likely intended to curb perceived censorship, but would also produce enormous legal headaches for tech companies that want to moderate harmful content for altruistic reasons. It is far from certain that a social media platform would win every case against it by proving that it consistently enforced its established standards. But even if a company managed to prevail every time, lawsuits can be brought for each piece of content removed, thus incurring enormous legal costs and serving as a significant drain on the company's time. The litigation costs would be high because the Florida law ensures that the stakes are high: Section 287.137 provides that a company that loses a case can be listed on a blacklist of antitrust violators, thus preventing it from doing business in Florida.[107]

Another problem is that the "consistent manner" aspect of content moderation stipulated by the law could be used as a legal hook by a wide variety of actors, including conspiracy theorists and disinformation operations propagandizing on behalf of foreign states. Such actors would have a cause of action that would at the very least burn up a tech company's time and resources when a legitimate outlet like CNN or the *New York Times* misreports breaking news, arguing that the company's treatment of CNN and the *Times* was not consistent with its treatment of the conspiracy theorists and disinformation operators.

There are other obvious tactics that the "consistent manner" provision gives rise to:

- Trolls could target prominent targets (e.g., people of color, LGBT people, religious minorities) with asymmetric harassment in the hope that they can provoke a comment made in the heat of the moment, under a storm of harassment, that is objectionable. Such a tactic is best carried out by anonymous accounts that act in unison with one another with the specific purposes of harassing/intimidating the target while trying to provoke the target to say something objectionable that gives rise to the argument that the target and the anonymous, harassing accounts should be subjected to the same treatment.

---

[106] "National Alliance: A Backgrounder," Anti-Defamation League, n.d., https://www.adl.org/resources/backgrounders/national-alliance-a-backgrounder.

[107] S.B. 7072, Sess. of 2021 (Fla. 2021), https://www.flsenate.gov/Session/Bill/2021/7072/BillText/er/HTML; Gov. Ron DeSantis, "Governor Ron DeSantis Signs Bill to Stop the Censorship of Floridians by Big Tech," May 24, 2021, https://www.flgov.com/2021/05/24/governor-ron-desantis-signs-bill-to-stop-the-censorship-of-floridians-by-big-tech/.

- The "consistent manner" provision also signals to trolls, harassers, and stalkers that the line that must be crossed for their activities to result in post deletion or user suspension will be rendered far more predictable. Thus, accounts acting in conjunction with one another could constantly probe to determine where the line lies. One account might say: *You should kill yourself*. The second says: *Have you thought about killing yourself?* Still another says: *You definitely shouldn't kill yourself*. The attacker waits to see which posts are pulled down and which accounts are banned. Immediately, the attacker knows how far harassment can be pushed. Further, the attacker can search for similar phraseology that can form the basis of a lawsuit claiming inconsistent moderation (for example, *you definitely shouldn't kill yourself* might be determined to be legitimate speech if said to someone who is suicidal; but the attacker can force the company to explain this to a court).

Essentially, the number of lines that a tech company must draw "consistently" are head spinning. New situations and events constantly emerge. Trying to be consistent across all situations, particularly in a manner that will satisfy a court, can be nearly impossible. The challenge would be accentuated under a law that, like Florida's, only allows new policies to be promulgated once every thirty days.

In essence, SB 7072 provides a huge disincentive for companies to engage in any but the most basic content moderation. This disincentive comes in the form of statutorily imposed inflexibility in content moderation (e.g., the prohibition on content moderation policy changes more than once every thirty days), increased ability for individuals to sue tech companies for their content moderation decisions, and statutory penalties designed to make the stakes high in this litigation. While we have shown that tech companies have not always drawn the lines right in their content moderation, this report also demonstrates that content moderation has served a positive social good in response to challenging situations, as a variety of malign actors have sought to exploit the power of social media. SB 7072 is designed to, and would have the effect of, making content moderation far more difficult in essentially all cases.

An Overview of Texas House Bill (HB) 20

Texas's HB 20 was introduced in the Texas House of Representatives as part of a special legislative session in August 2021. The bill claims to "protect Texans from wrongful censorship on social media platforms," primarily by preventing social media companies with over 50 million users from banning content and users based on their political viewpoint.[108] Before the bill's passage, Gov. Greg Abbott identified social media censorship as a key theme of the special legislative session.[109] HB 20 was sponsored by a coalition of 65 Republican state senators and representatives.[110] Gov. Abbott signed the controversial bill into law on September 9, 2021. Similar to SB 7072, HB 20 was subjected to intense debate in the legislature and in the media. Indeed, many political opponents and industry

[108] Office of the Texas Governor, "Governor Abbott Signs Law Protecting Texans From Wrongful Social Media Censorship," September 9, 2021, https://gov.texas.gov/news/post/governor-abbott-signs-law-protecting-texans-from-wrongful-social-media-censorship.

[109] Drew Knight, "Texas Social Media Censorship Bill Signed into Law," *KVUE*, September 9, 2021, https://www.kvue.com/article/news/politics/texas-social-media-censhorship-bill-signed-house-bill-20/269-54c04a75-2bc1-419b-928f-d5d8c39884f4.

[110] "TX HB20 | 2021 | 87th Legislature 2nd Special Session," *LegiScan*, September 9, 2021, https://legiscan.com/TX/bill/HB20/2021/X2.

analysts compared it to the Florida bill.[111] On September 22, 2021, the same technology industry associations that filed suit against SB 7072, NetChoice and CCIA, did so against HB 20 in the U.S. District Court for the Western District of Texas, citing similar First Amendment and Section 230 concerns. The court issued a preliminary injunction on December 1, 2021, which the State of Texas appealed. Though the Fifth Circuit lifted the injunction, the Supreme Court subsequently reinstated it. As mentioned earlier, in September 2022 the Fifth Circuit found that HB 20 was constitutional and authorized its implementation, setting up a likely Supreme Court challenge.[112]

Texas's approach to its anti-censorship legislation is somewhat less punitive than is Florida's law, in that HB 20 contains no provisions that could result in the blacklisting of social media companies from doing business in the state. However, HB 20 on the whole covers more ground than does SB 7072.

Indeed, HB 20 begins from the premise that social media platforms are common carriers, stating that "social media platforms function as common carriers, are affected with a public interest, are central public forums for public debate, and have enjoyed governmental support in the United States; and social media platforms with the largest number of users are common carriers by virtue of their market dominance."[113] This provision could potentially become a thorny legal issue down the road. Courts have to some extent have already weighed in on the issue of what constitutes a common carrier in the information space. One federal court ruling held that companies like Facebook, Google, Twitter, and YouTube "are not considered common carriers that hold themselves out as affording neutral, indiscriminate access to their platform without any editorial filtering."[114] The Supreme Court has also held that, even where a service has been deemed a common carrier, the First Amendment protects it from compelled speech (holding that California could not require a utility company to include a third party's newsletters in bills that it sent to customers).[115] These legal precedents led the district court reviewing HB 20 to reject the claim outright, determining that social media platforms are *not* common carriers.[116] Yet it is not difficult to imagine such a point being debated for years to come, with different courts and commentators potentially reaching radically different conclusions.

Section 120.052 of HB 20 requires the creation of an acceptable use policy that 1) stipulates the types of content allowed on a platform, 2) describes how the platform ensures compliance with its

---

[111] See, for example, Robert F. Kennedy Human Rights, Press Release, "Texas: Why We Oppose HB 20," April 27, 2021, https://rfkhumanrights.org/press/texas-why-we-oppose-hb-20; Steve DelBianco, "RE: Opposing HB 20- Relating to Complaint Procedures and Disclosure Requirements for, and to the Censorship of Users' Expressions by, Social Media Platforms," House Select Committee on Constitutional Rights & Remedies, Texas House of Representatives, August 22, 2021 https://netchoice.org/wp-content/uploads/2021/08/NetChoice-Opposition-to-Texas-HB-20-Special-Session-2.pdf; Kailyn Rhone, "Social Media Companies Can't Ban Texans Over Political Viewpoints Under New Law," *The Texas Tribune*, September 2, 2021, https://www.texastribune.org/2021/09/02/texas-social-media-censorship-legislature/; Nathan Sheard, "Texas' Social Media Law is Not the Solution to Censorship," Electronic Frontier Foundation, September 15, 2021, https://www.eff.org/deeplinks/2021/09/texas-social-media-law-not-solution-censorship.

[112] *NetChoice v. Paxton*, No. 21-51178 (5th Cir., September 16, 2022), p. 2, https://www.ca5.uscourts.gov/opinions/pub/21/21-51178-CV1.pdf.

[113] H.B. 20, Sess. of 2021 (Tex. 2021), https://capitol.texas.gov/tlodocs/872/billtext/html/HB00020F.htm.

[114] *United States Telecom Association v. FCC*, 855 F.3d 381, 392 (D.C. Cir., 2017).

[115] *Pacific Gas & Electric Co. v. Public Utilities Commission,* 475 U.S. 1 (1986).

[116] *NetChoice v. Paxton*, Order, 1:21-CV-840-RP (W.D. Tex., December 1, 2021), p. 15, https://s3.documentcloud.org/documents/21124083/govuscourtstxwd1147630510.pdf.

policy, 3) explains how users can report illegal content or content that violates this policy, and 4) requires a biannual report detailing actions taken to enforce the policy. Some of these elements are similar to the terms of service or end-user license agreements already in place for major social media platforms. For example, Twitter's Rules and Policies page features dozens of links detailing Twitter's policies regarding use of its platform.[117] Facebook also maintains a similar site with dozens of links to specific policies, including policy rationale.[118] Google and YouTube also feature similar pages.[119] The biannual transparency report stipulated in HB 20 is intended to present data across the preceding six-month period on a variety of metrics related to content takedowns. Similar to the acceptable-use policies that most social media companies have in place, platforms like Twitter, Facebook, and YouTube already produce transparency reports, some of which are published on a quarterly basis.[120] However, these policies and reports are currently not subject to legal oversight and potentially a legal process of discovery. HB 20 would change that.

The remainder of Section 120 lays out other compliance elements that social media companies must implement. This includes establishing a complaint system whereby users can submit complaints about illegal content or complain about removal of content posted by the user. Companies must make a good-faith effort to process these complaints within 48 hours.[121] If a company removes content that violates its platform's acceptable use policy, the company must notify the user why the content was removed, provide an appeal option and appeal follow-up, as well as a reversal notification, if applicable (though these requirements do not apply if the platform removes content pursuant to a law enforcement investigation).

Section 143A.002, entitled "Censorship Prohibited," serves as perhaps the core component of HB 20, and is much broader in scope than is the anti-censorship provision in SB 7072. The section states:

> A social media platform may not censor a user, a user's expression, or a user's ability to receive the expression of another person based on: the viewpoint of the user or another person; the viewpoint represented in the user's expression or another person's expression; or a user's geographic location in this state or any part of this state.
>
> This section applies regardless of whether the viewpoint is expressed on a social media platform or through any other medium.[122]

---

[117] Twitter, "Rules and Policies," 2022, https://help.twitter.com/en/rules-and-policies#platform-use-guidelines.
[118] Meta Transparency Center, "Policies," 2022, https://transparency.fb.com/policies/.
[119] Google, "Privacy & Terms," 2022, https://policies.google.com/terms?hl=en-US#toc-removing; YouTube, "Terms of Service," 2022, https://www.youtube.com/static?template=terms.
[120] See, for example, Twitter Transparency Center, "Rules Enforcement Jan - Jun 2021," January 25, 2022, https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jan-jun; Meta Transparency Center, "Community Standards Enforcement Report Q3 2021," November 2021, https://transparency.fb.com/data/community-standards-enforcement/; Google, "YouTube Community Guidelines Enforcement," 2021, https://transparencyreport.google.com/youtube-policy/removals.
[121] H.B. 20, Sec. 120.101, Sess. of 2021 (Tex. 2021), https://capitol.texas.gov/tlodocs/872/billtext/html/HB00020F.htm.
[122] Ibid., Sec. 143A.002.

Where SB 7072 focused on consistency of content moderation, HB 20 adopts a broader approach by affirmatively preventing "censorship" based on the viewpoint presented in the content.

Two subsections include important information about enforcement of the law. Sec. 143A.005 acknowledges that "this chapter does not subject a social media platform to damages or other legal remedies to the extent the social media platform is protected from those remedies under federal law." Determining the intersection between HB 20 and Section 230 could potentially lead to murky legal questions in the future. For example, Sec. 143A.006(a) includes several carve-outs that allow platforms to restrict expression that:

1) The social media platform is specifically authorized to censor by federal law;

2) Is the subject of a referral or request from an organization with the purpose of preventing the sexual exploitation of children and protecting survivors of sexual abuse from ongoing harassment;

3) Directly incites criminal activity or consists of specific threats of violence targeted against a person or group because of their race, color, disability, religion, national origin or ancestry, age, sex, or status as a peace officer or judge; or

4) Is unlawful expression.[123]

Like SB 7072, HB 20 allows users to take legal action against social media companies for violations of the law. Aggrieved users can seek either declaratory relief (including costs and reasonable attorney fees) or injunctive relief. Additionally, the Texas Attorney General can also receive reports from users about alleged violations of the law and may file for injunctive relief as a result of an investigation.

Considerations Based on the Law's Text

HB 20 is intended to "protect Texans from wrongful social media censorship" by restricting social media companies' ability to moderate content based on its political viewpoint. However, the law's imprecision, loopholes, and other shortcomings allow exploitation by malign actors. What follows are examples of the ways that HB 20 could be used by such individuals and entities.

Many of the same loopholes we discussed with respect to SB 7072, particularly those related to SB 7072's "consistent manner" provision, apply to HB 20. And, like SB 7072, HB 20 reveals companies' content moderation playbooks to everyone, including bad actors. Also applicable are the concerns we raised about asymmetric harassment and testing the limits of how far bad behavior can extend through the use of fake accounts. In essence, HB 20 generates the same kind of inflexible system for responding to bad actors' use of social media platforms that would be created by SB 7072.

Yet HB 20 also has some unique loopholes. HB 20's ban on removal of content based on viewpoint limits the ability of a platform to remove toxic material, potentially opening the door for large amounts of highly objectionable content to be prohibited by law from being removed by companies. For example, does the takedown of pro-ISIS content constitute censorship on the basis of

[123] Ibid., Sec. 143A.006(a).

viewpoint? While ISIS itself is a proscribed terrorist organization, posts that are objectively pro-ISIS are, in most cases, perfectly legal. While recruiting for the group, fundraising for the group, or saying "come fight for the caliphate" might be illegal, under U.S. law proclaiming love for ISIS in a tweet is not, nor is admiring the group or celebrating its successes. At its peak, ISIS had tens of thousands of sympathizers, many of whom openly expressed their enthusiasm for the group. Many of these individuals were not even ISIS members. Under HB 20, there is a strong legal argument that their pro-ISIS content can no longer be moderated. Isn't any violent extremist cause nothing more than a viewpoint in the absence of illegal activity? If this is so, any violent extremist material could be protected from moderation under HB 20. The same concern applies to hate speech appearing on a platform. Don't discriminatory statements, under the Texas law, just constitute "viewpoints"?

HB 20 includes provisions that appear intended to address this issue. But they fail to fully do so. For example, one may read Sec. 143A.006(a)(3), quoted above, as an anti-hate speech provision, as it allows companies to remove speech that "consists of specific threats of violence targeted against a person or group because of their race, color, disability, religion, national origin or ancestry, age, sex, or status as a peace officer or judge." However, because the provision is crafted so narrowly, it fails to allow moderation of most kinds of hate speech. For companies to be allowed to remove hate speech under HB20's regime, it must consist of "*specific threats* of violence" against a person or group based on the enumerated factors. The fact is, most online hate speech and harassment does not contain explicit threats of violence, let alone specific ones.

Current Status of the SB 7072 and HB 20 – Legal Challenges

Given the controversy surrounding SB 7072 and HB 20 since their conception, observers anticipated legal challenges. Indeed, both laws were enjoined by federal courts just before they were scheduled to go into effect. This section examines the legal issues in these federal cases.

SB 7072. On June 30, 2021, SB 7072 was enjoined by the U.S. District Court for the Northern District of Florida. The preliminary injunction was issued in response to a lawsuit brought by NetChoice and CCIA. Part of the court's justification for issuing a preliminary injunction stemmed from its finding that the plaintiffs were likely to succeed in their claim that SB 7072 was preempted by Section 230 of the Communications Decency Act.

The court also held that restricting platforms' editorial discretion posed First Amendment issues. The court first rejected Florida's argument that it was the state, and not the plaintiffs, who were on the side of the First Amendment, describing this contention as "perhaps a nice sound bite" that is in fact "wholly at odds with accepted constitutional principles." The court explained:

> The First Amendment says "Congress" shall make no law abridging the freedom of speech or of the press. The Fourteenth Amendment extended this prohibition to state and local governments. The First Amendment does not restrict the rights of private entities not performing traditional, exclusive public functions. See, e.g., *Manhattan Cmty. Access Corp. v. Halleck*, 139 S. Ct. 1921, 1930 (2019). So whatever else may be said of the providers' actions, they do not violate the First Amendment.[124]

[124] Preliminary Injunction, *NetChoice LLC v. Moody*, 4:21-cv-00220-RH-MAF (N.D. Fla., June 30, 2022), p. 17.

The court then explained that the law's explicit purpose was "reining in the ideology of the large social-media providers," which is "precisely the kind of state action held unconstitutional" in previous First Amendment cases.[125] SB 7072, the court wrote, "compels providers to host speech that violates their standards—speech they otherwise would not host—and forbids providers from speaking as they otherwise would."[126]

The court also took issue with SB 7072's requirement that content be moderated in a consistent manner:

> The statute does not define "consistent manner." And the statute does not address what a social media platform should do when the statute itself prohibits consistent application of the platform's standards—for example, when a candidate engages in conduct that would appropriately lead to deplatforming any other person, or when content "by or about" a candidate, if by or about anyone else, would be post-prioritized, or when a "journalistic enterprise" posts content that would otherwise be censored.[127]

Immediately after the district court issued a preliminary injunction, Florida filed an appeal with the U.S. Court of Appeals for the Eleventh Circuit, which upheld the District Court's decision.

HB 20. On December 1, 2021, one day before the law was set to go into effect, HB 20 was enjoined by the U.S. District Court for the Western District of Texas. As was the case for SB 7072, this preliminary injunction came in response to a lawsuit brought by NetChoice and CCIA.

The court's decision to enjoin HB 20 centered on the constitutional issues raised by the law—namely First Amendment issues stemming from the law's limitations on platforms' editorial discretion: HB20 forces companies to host speech that is not in line with the type of community they are attempting to create. The court specifically singled out the way that hate speech would likely be defined as viewpoint-based under HB 20, and thus protected from "censorship" by social media platforms:

> The threat of lawsuits for violating Section 7 of HB 20 chills the social media platforms' speech rights. HB 20 broadly prohibits content moderation based on "viewpoint," authorizing the Texas Attorney General to sue for violations—and even "potential" violations—of Section 7's "censorship" restrictions. In response to the State's interrogatories, NetChoice explained that the "threat of myriad lawsuits based on individual examples of content moderation threaten and chill the broad application of those [content moderation] policies, and thus H.B. 20's anti-moderation provisions interfere with Plaintiff's members' policies and practices.... Using YouTube as an example, hate speech is necessarily 'viewpoint'-based, as abhorrent as those viewpoints may be. And removing such hate speech and assessing

[125] Ibid., p. 21.
[126] Ibid., p. 1.
[127] Ibid., pp. 11-12.

> penalties against users for submitting that content is 'censor[ship]' as defined by H.B. 20."[128]

The court rejected the notion that social media platforms are common carriers and took issue with the "inordinately burdensome" requirements of HB 20, which require appeals on all content removals. To illustrate the burden, the court detailed the sheer volume of material that would immediately require an appeals process under HB 20:

> Section 2's disclosure and operational provisions are inordinately burdensome given the unfathomably large numbers of posts on these sites and apps. For example, in three months in 2021, Facebook removed 8.8 million pieces of "bullying and harassment content," 9.8 million pieces of "organized hate content," and 25.2 million pieces of "hate speech content." During the last three months of 2020, YouTube removed just over 2 million channels and over 9 million videos because they violated its policies. While some of those removals are subject to an existing appeals process, many removals are not. For example, in a three-month period in 2021, YouTube removed 1.16 billion comments. Those 1.16 billion removals were not appealable, but, under HB 20, they would have to be. Over the span of six months in 2018, Facebook, Google, and Twitter took action on over 5 billion accounts or user submissions—including 3 billion cases of spam, 57 million cases of pornography, 17 million cases of content regarding child safety, and 12 million cases of extremism, hate speech, and terrorist speech. During the State's deposition of Neil Christopher Potts ("Potts"), who is Facebook's Vice President of Trust and Safety Policy, Potts stated that it would be "impossible" for Facebook "to comply with anything by December 1, [2021]. . . [W]e would not be able to change systems in that nature. . . . I don't see a way that we would actually be able to go forward with compliance in a meaningful way."[129]

While HB 20 contained a severability clause, the court ruled that this clause did not preclude the law's facial invalidation. The court held that both sections 2 and 7 of HB 20 were "replete with constitutional defects, including unconstitutional content- and speaker-based infringement on editorial discretion and onerously burdensome disclosure and operational requirements." Thus, the court found that, "like the Florida statute, '[t]here is nothing that could be severed and survive.'"[130]

Immediately after the district court issued its preliminary injunction, the State of Texas filed an appeal with the Fifth Circuit Court of Appeals.[131] Almost a year and half later, a split 2-1 panel of Fifth Circuit judges lifted the district court's injunction without ruling on the merits of the original lawsuit, leaving the constitutional issues still in play.[132] We lack insight into the majority's reasoning in its May 11, 2022 one-sentence decision because no written opinion was issued to accompany it.

---

[128] Order, *NetChoice, LLC v. Paxton*, 1:21-CV-840-RP (W.D. Tex., December 1, 2021), p. 19, https://s3.documentcloud.org/documents/21124083/govuscourtstxwd1147630510.pdf.
[129] Id., pp. 21-22 (citations removed).
[130] Id., pp. 28-29.
[131] John Villasenor, "Texas' New Social Media Law Is Blocked for Now, But That's Not the End of the Story," *The Brookings Institution*, December 14, 2021, https://www.brookings.edu/blog/techtank/2021/12/14/texas-new-social-media-law-is-blocked-for-now-but-thats-not-the-end-of-the-story/.
[132] Order, *NetChoice v. Paxton*, No. 21-51178 (5th Cir., May 11, 2022), p. 1, https://netchoice.org/wp-content/uploads/2022/05/2022-05-11-Court-Order-dckt-.pdf.

Two days later, NetChoice and CCIA submitted an emergency request to the Supreme Court asking the Court to reinstate the district court ruling that blocked HB 20 from taking effect.[133] The Supreme Court reversed the Fifth Circuit, putting the injunction back into effect.[134] The Fifth Circuit subsequently issued a decision on the merits, ruling in favor of the State of Texas and authorizing the implementation of HB 20.

On September 21, 2022, five days after the Fifth Circuit issued its ruling that content moderation was indeed constitutional, the State of Florida submitted a petition for writ of certiorari to the U.S. Supreme Court, asking the Court to answer two questions. In its petition, Florida frames the two key questions:

1. Whether the First Amendment prohibits a State from requiring that social media companies host third party communications, and from regulating the time, place, and manner in which they do so.

2. Whether the First Amendment prohibits a State from requiring social media companies to notify and provide an explanation to their users when they censor the user's speech. [135]

Other Anti-Censorship Legislative Efforts

Aside from SB 7072 and HB 20, there have been multiple other state and federal legislative initiatives targeting social media companies' content moderation policies. "Anti-censorship" efforts are likely to be a recurring political issue in state legislatures as well as the US Congress. We now provide a brief examination of efforts outside of Texas and Florida to alter the content moderation status quo.

Mere days after President Trump's expulsion from major social media platforms in January 2021, a slew of states introduced bills similar to SB 7072 and HB 20. One list assembled by research analyst Megan Kashtan at the private firm Leadership Connect documents over a dozen pieces of legislation with an "anti-censorship" bent that are designed to regulate major social media platforms. Legislators in Alabama, Arkansas, North and South Dakota, Idaho, Iowa, Kansas, Kentucky, Louisiana, Missouri, North Carolina, Oklahoma, Utah, West Virginia, and Wyoming have introduced bills on the matter.[136] However, only Texas and Florida have successfully passed anti-censorship bills into law. The other pieces of legislation in the predominantly Republican states where they were introduced have either not yet made it out of the legislature or have been vetoed by their governors (with the governor of Utah vetoing SB 228).

---

[133] See Amy Howe, "Tech Industry Asks Court to Block Texas Law that Targets Social Media Companies," *SCOTUS Blog*, May 14, 2022, https://www.scotusblog.com/2022/05/tech-industry-asks-court-to-block-texas-law-that-targets-social-media-companies/.

[134] Order, *NetChoice v. Paxton*, No. 21-51178 (5th Cir., May 11, 2022), p. 1, https://netchoice.org/wp-content/uploads/2022/05/2022-05-11-Court-Order-dckt-.pdf.

[135] Petition for Writ of Certiorari, Moody v. NetChoice, No. (11th Cir., September 21, 2022), https://netchoice.org/wp-content/uploads/2022/09/Moody-FL-v.-NetChoice-Petition-for-Certiorari-filed-Sept-21-2022.pdf.

[136] Megan Kashtan, "Tracking Proposed Social Media Legislation in America," *Leadership Connect*, April 29, 2021, https://www.leadershipconnect.io/business/2021/04/29/tracking-proposed-social-media-legislation-in-america/.

A similar flurry of activity targeting social media platforms can be seen at the federal level in recent years regarding Section 230. As this report details, tech companies have become increasingly controversial, particularly on the political right. President Trump thus explored ways to repeal Section 230, including through the use of executive orders or by the FCC.[137] A list published by *Slate* tracking legislation targeting Section 230 by either repealing it, limiting its scope, introducing new obligations to qualify for liability protection, or altering the "Good Samaritan" provision shows 20 bills introduced in the 117th Congressional Session thus far.[138]

For example, Sen. Lindsay Graham (R-S.C.) has introduced a bill designed to repeal Section 230 in its entirety. Also of note is the Preserving Political Speech Online Act, introduced by Sen. Steve Daines (R-MT), which is designed to limit the Good Samaritan provision of Section 230. *Slate* explains: "Currently, platforms receive Section 230 protections only when they remove content 'in good faith' that they consider to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable. Under the Preserving Political Speech Online Act, acceptable reasons for 'good faith' removal would be limited to content that is obscene, illegal, or excessively violent."[139] And Rep. Jim Jordan's (R-OH) Protect Speech Act seeks to limit the scope of liability protection provided by Section 230. *Slate* explains:

> To use Section 230 as a defense, the platform would need to publicly state terms of service that detail criteria used in content moderation decisions. Platforms would also need to comply with those stated terms of service and content moderation criteria, and would need to ensure that content moderation is not made on deceptive grounds. When content is restricted, platforms would need to provide a rationale and an opportunity for the user to respond, with certain exceptions for law enforcement and imminent threats to safety.[140]

There have also been numerous other pieces of federal legislation introduced from both sides of the aisle. *Slate*'s roundup of the various ways that Members of Congress want to change Section 230 is worth reading for anyone interested in the likely future of this political battle.

Outlook: The Dark Side of Anti-Censorship Legislation

The national debate about content moderation is fierce because fundamental principles are at stake: freedom of speech and political expression, personal safety, and the role of major corporations in regulating platforms that have become central to the way many Americans and people worldwide communicate their ideas and with one another. It is outside the purview of this report to make recommendations for how social media companies can improve content moderation efforts, but we acknowledge that there are legitimate concerns about how they have applied their policies in practice.

Decisions about content removal and deplatforming can often appear inconsistent and arbitrary to

[137] See discussion in Sara Morrison, "How the Capitol Riot Revived Calls to Reform Section 230," *Vox*, January 11, 2021, https://www.vox.com/recode/22221135/capitol-riot-section-230-twitter-hawley-democrats.
[138] Meghan Anand et al., "All the Ways Congress Wants to Change Section 230," *Slate*, March 23, 2021, https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html.
[139] Ibid.
[140] Ibid.

outsiders. The confusion around moderation standards has undoubtedly reinforced allegations of bias, allegations that in some cases may have a basis in fact. This perception has led to the passage of laws in Florida and Texas that attempt to diminish bias in content moderation, raising the bar for removal of content, and providing legal redress for those damaged by social media company actions. The upshot, however, is a major discrepancy between the legitimate intent to enhance objective decisionmaking and the public interest.

The recent anti-censorship laws are ill-advised because they shackle social media companies from responding adroitly to the evolving tactics of malign actors and create several large loopholes that such actors can readily exploit. Content moderation playbooks are far from perfect, but they are continually being refined. Overly restrictive policies on regulating licit content on social media, which vastly outweighs illicit content, tilts the advantage to those who are bent on harming people with their words and, as is too often the case, also with their actions. We know that what happens on social media does not stay on social media. The risk of real-world anguish, trauma, injury, or even violence from malicious content on the internet is high.

Recent anti-censorship legislation is flawed because in erecting new protections for users of social media and enforcing stricter standard-setting it is ripe for abuse. Under SB 7072's provision that protects journalistic enterprises, Florida could see a proliferation of media sites claiming to be journalistic enterprises that are in fact aimed solely at propagating violent extremist ideologies, espousing hateful ideas, or advancing a U.S. adversary's objectives. Likewise, the political candidate exception in SB 7072 could be milked by malign actors who, by running for election in Florida, can ensure they are able to maintain an online platform for their violent extremist views, hate speech, or mis/disinformation.

The requirements in SB 7072 and HB 20 for greater publication of standards and the possibility of legal action to compel the release of specifics is ripe for abuse. They could create an iterative process whereby an individual or group continually tests the limits of moderation and content removal, and can innovate based on the discovery of company restrictions. A foreign influence or harassment campaign, as well as propagators of content that advocates for but does not incite violence, could all make use of this technique.

Consistency in content moderation and lack of discrimination based on viewpoint are legitimate goals, but the application of these principles in the Florida and Texas laws is problematic. It is unclear how SB 7072's stringent consistency requirement would impact asymmetric harassment or the disaggregation of conspiracy theories from hard news reporting, especially when new rule promulgation is limited by the law to once a month. Further, the ease with which litigants can claim inconsistency in moderation virtually ensures that litigation would increase considerably.

HB 20's viewpoint protections are even more expansive than anything in the Florida law. Almost all forms of extremism, hatred, and even mis/disinformation constitute a viewpoint. Expressing a favorable view of racial supremacy or disseminating misleading information about public health issues is not inherently illegal. Under HB 20, platforms would be hamstrung in removing posts that propagate such "viewpoints."

Efforts to reform content moderation are playing out in a highly uncertain legal environment. Pressure for change at the federal and state levels is likely to persist and possibly intensify. Those who are interested in sound approaches to this thorny issue should not count on existing federal law

to protect against overreach by anti-censorship proponents. Further, as it stands, stipulations in anti-censorship laws that they cannot run afoul of federal law is flimsy assurance against abuse. Enforcement of any anti-censorship legislation through court processes will not be clean and is certain to be mired in the complexities of intersecting laws and murky judicial interpretations.

Today we are a far cry from the early days of content moderation when everyone agreed on the national security justification for clamping down on ISIS's aggressive online activities. Dangerous if often murky actors have proliferated online while the politics of social media have become increasingly divisive. Ensuring that social media spaces do not become safe havens for those who seek to spread toxic content—rather than just share provocative opinions—is a reasonable goal. Anti-censorship legislation as currently conceived will not help social media companies draw content moderation lines more carefully, balancing freedom of speech with personal safety. Rather, such laws will be a boon to extremists and other malign actors, allowing them to bust open the floodgates that have barely kept them contained.