Disrupting the Narrative: Diving Deeper into Section 230 Political Discourse

Authors:

Koustubh "K.J." Bagchi[a]

Elizabeth Banker[b]

Ife Ogunleye[c]

November 2022

Working Paper

The Center for Growth and Opportunity at Utah State University is a university-based academic research center that explores the scientific foundations of the interaction between individuals, business, and government.

[a] Koustubh "K.J." Bagchi is Senior Director of Technology Policy at the Chamber of Progress and the author of several influential pieces on Section 230. In addition to advising members of the Washington State Senate, he worked as Legislative Counsel for a D.C. City Councilmember and former Rep. Mike Honda (D-CA), a Member of Congress who served on the influential House Appropriations Committee.

[b] Elizabeth Banker is the Vice President of Legal Advocacy for the Chamber of Progress. Banker is an adjunct professor of law at UC College of the Law, San Francisco and previously taught as an adjunct professor of law at Georgetown Law Center.

[c] Ife Ogunleye is the Manager for Technology and Human Rights in the Americas at BSR. Ogunleye's policy experience covers anti-corruption, artificial intelligence, privacy, and tech. Ife has a Master of Development Practice from the University of California, Berkeley and a law degree from the University of Manchester, England.

The Center for Growth and Opportunity at Utah State University

The Dark Side of the Web

During the past few decades, the proliferation of online spaces that allow for discourse and the exchange of digital information has highlighted, for many of us, the darker corners of our society. For example, in the past two years, the spread of COVID-19 not only saw countries combating the spread of a virus but also communities latching onto vaccine misinformation online. Further, hate speech continues to exist on digital platforms despite the innovative AI tactics employed by companies to combat it and the countless resources dedicated to hiring and training human content reviewers. Moreover, people who participate in the scourge of child exploitation online have used online spaces to produce and spread their horrifying content.

Federal policymakers have introduced a variety of bills aimed at curbing all three of the harmful trends described above, but the focus of this paper will be on the narrative and content related to legislation aimed at revoking or limiting the application of Section 230 of the Communications Decency Act. More importantly, we will cover in depth what actions platforms are undertaking in these areas to combat some of the challenges raised by open online discourse. Finally, we argue that Section 230 plays an important role in promoting provider efforts to curb the types of online content that these legislative proposals seek to address.

Section 230, passed as part of the Communications Decency Act of 1996,[1] has drawn criticism for the protections it gives providers who post or host third-party content online. The law acts as a shield against lawsuits that seek to treat providers as though they are publishers or distributors of the online content available through their services, so long as what is illegal about the content originates with another party.[2] It also provides important protections for providers when they remove or reject content that violates their rules.[3] In essence, Section 230 bolsters the freedom-of-speech protections granted by the First Amendment for social media providers.[4]

Section 230 promotes content moderation by preventing the application of traditional publisher-distributor liability laws to online providers. In the absence of Section 230, providers who engage in content moderation would be subject to lawsuits and potential liability for content created by their users in the same way that a newspaper is liable for the content written by its employees.[5] At the same time, providers who do not moderate content face less legal risk, because there are constitutional limitations on imposing strict liability on "secondary publishers" or distributors for content created by others.[6]

Thus, the application of traditional publisher-distributor liability creates what is called the "moderator's dilemma": a regime that protects from liability, providers who take a hands-off approach to content on their services while making the providers who enforce content rules subject

1 Pub. L. No. 104-104, 110 Stat. 56, 137–39 (1996).

2 47 U.S.C. § 230(c)(1) (1996) The Ninth Circuit Court of Appeals has interpreted Section 230 to apply unless providers "materially contributed to illegality" of content. Fair v. Roommates, 521 F.3d 1157, 1167–68 (9th Cir. 2008).

3 47 U.S.C. § 230(c)(1), (c)(2); Barnes v. Yahoo!, Inc., 565 F.3d 560, 569 (9th Cir. 2009). The court held that "Subsection (c)(1), by itself, shields from liability all publication decisions, whether to edit, to remove, or to post, with respect to content generated entirely by third parties. Subsection (c)(2), for its part, provides an additional shield from liability, but only for 'any action voluntarily taken in good faith to restrict access to or availability of material that the provider . . . considered to be obscene . . . or otherwise objectionable.'"

4 Reese D. Bastian, "Content Moderation Issues Online: Section 230 Is Not to Blame," *Texas A&M Journal of Property Law* 8 (2022): 43–72, https://doi.org/10.37419/JPL.V8.I2.1.

5 *Barnes*, 565 F.3d at 568.

6 Smith v. California, 361 U.S. 147, 152–53, 80 S. Ct. 215, 218–19, 4 L.Ed.2d 205 (1959).

to lawsuits and liability risks.[7] Without Section 230, there would be substantial risks to online providers who take actions to moderate content such as those described below. Thus, removing Section 230 protections is likely to reduce the level of content moderation that is available on online platforms—the opposite of what the bills discussed seek to promote—and in fact to make the problems of content moderation worse.[8]

We examine how Section 230 enables platforms to conduct content moderation concerning health misinformation, hate speech, and child sexual abuse materials below.

What does Section 230 of the Communications Decency Act Do?

Section 230 of the Communications Decency Act of 1996 provides companies with protection against liability for user-generated content on social media platforms and online services. The law states that companies are not liable for content published by others and for actions the company takes to restrict access to materials it considers objectionable. In particular, Section 230(c) 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" and "no provider or user of an interactive computer service shall be held liable on account of any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[9]

In addition to protecting companies from liability for third-party content, Section 230 also provides procedural benefits to companies by ensuring that litigation cases brought against them on account of a content moderation decision (i.e., the decision to remove or maintain content uploaded by a third-party user) can be dismissed in the early stages, thereby mitigating the heavy costs and damages that lawsuits can create for businesses. These protections have come in handy for companies that seek to moderate illegal, hateful content or misinformation, and they have been used as a defense against suits brought by dissatisfied users or third parties.

In recent times, however, Section 230 has been criticized by lawmakers and policymakers from both sides of the political divide as the reason for the lack of content moderation, the reason for the spread of misinformation and disinformation, and the tool by which online platforms are excluding conservative speech and punishing those who hold opposing political views. In reality, the protections provided by Section 230 have empowered companies to safely carry out content moderation and host a wide range of content and speech on their platforms.

7 Zeran v. America Online, Inc., 129 F.3d 327 (4th Cir. 1997) @ 331.

8 See Jeff Kosseff, "A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not)," *Berkeley Technology Law Journal* 37, no. 2 (2022): 31, https://ssrn.com/abstract=3905347.

9 Pub. L. No. 104-104, 110 Stat. 56, 138 (1996).

Health Misinformation

Misinformation has become a major issue in public policy. Various studies have found that Americans consider misinformation to be a problem that needs to be addressed. In a 2021 poll, approximately 95 percent of Americans identified the spread of misinformation as a problem: 81 percent of respondents viewed it as a major problem and 13 percent viewed it as a minor problem.[10] Seventy-five percent of the respondents also indicated that they worry about their exposure to misinformation and about the potential that their friends or family have spread misinformation, even unintentionally.[11] A large majority of respondents to this poll also considered users of social media and social media companies responsible for addressing the problem.[12] An online survey by the Chamber of Progress also found that when voters were asked about policy issues that should be focused on, two of the top four priorities identified were combating violent extremists and stopping COVID-19 misinformation online.[13] It is therefore important to examine the state of misinformation online and assess what steps companies are currently taking to address it.

On July 22, 2022, Senators Amy Klobuchar (D-MN) and Ben Ray Luján (D-NM) introduced the Health Misinformation Act. In the bill's press release, the senators argue that Section 230—"which was intended to promote online speech and allow online services to grow—now distorts legal incentives for platforms to respond to digital misinformation on critical health issues, like Covid-19, and leaves people who suffer harm with little to no recourse."[14] According to the bill's original sponsors, "The Health Misinformation Act would create an exception to Section 230 of the Communications Decency Act's liability shield for platforms with algorithms that promote health-related misinformation related to an existing public health emergency, as declared by the Secretary of Health and Human Services (HHS). The legislation directs HHS to issue guidelines as to what constitutes health misinformation."[15]

Since early 2020, when COVID-19 was declared a global pandemic, major platforms have developed and executed content moderation policies regarding COVID-19 misinformation and disinformation.[16] These policies were later updated to address misinformation and disinformation about the vaccines.[17] Reviews of the community guidelines and policies updated and created in the wake of the pandemic show that most social media platforms had developed a range of remedies to address misinformation, including soft measures, such as warnings, and hard measures, such as removals.[18]

---

10 Pearson Institute and AP-NORC Center for Public Affairs Research, "The American Public Views the Spread of Misinformation as a Major Problem," 2021, https://apnorc.org/wp-content/uploads/2021/10/misinformation_Formatted_v2-002.pdf.

11 Pearson and AP-NORC, "American Public Views the Spread of Misinformation."

12 Pearson and AP-NORC, "American Public Views the Spread of Misinformation."

13 Chamber of Progress, "Survey of National Voters," 2022, http://progresschamber.org/wp-content/uploads/2022/04/NPA3462-COP-National-FINAL.pdf.

14 Amy Klobuchar, "Klobuchar, Luján Introduce Legislation to Hold Digital Platforms Accountable for Vaccine and Other Health-Related Misinformation," press release, July 22, 2021, https://www.klobuchar.senate.gov/public/index.cfm/2021/7/klobuchar-luj-n-introduce-legislation-to-hold-digital-platforms-accountable-for-vaccine-and-other-health-related-misinformation.

15 Klobuchar, "Klobuchar, Luján Introduce Legislation."

16 Spandana Singh and Koustubh "K.J." Bagchi, "How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19," Open Technology Institute, June 1, 2020, https://www.newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/.

17 Nandita Krishnan et al., "Research Note: Examining How Various Social Media Platforms Have Responded to COVID-19 Misinformation," *Harvard Kennedy School Misinformation Review* 2 (2021), https://misinforeview.hks.harvard.edu/article/research-note-examining-how-various-social-media-platforms-have-responded-to-covid-19-misinformation/.

18 Krishnan et al., "Research Note."

For example, since September 2021, Google's misinformation policy has included vaccine misinformation.[19] Twitter upped the consequences for repeat violations of its COVID-19 misinformation policy by implementing a 12-hour account lock for two strikes and a permanent account suspension for five or more strikes.[20] Finally, Meta extended its misinformation policy to COVID-19 misinformation in 2020. It removes posts that contain misinformation on the basis of guidance from health authorities and works with fact-checking partners to debunk conspiracy theories and other misleading theories.[21] In addition to implementing internal policy updates, companies also worked with governments and health agencies to elevate authoritative sources related to COVID-19 by directing users to reliable information from official sources when the users logged on or searched for related content.[22]

All three major platforms reported on the impact of their implemented policies. As of 2021, Google reported removing 130,000 videos from YouTube for violating COVID-19 vaccine misinformation policies.[23] Between February 2020 and August 2021, more than 1 million videos with misinformation related to COVID-19 (including claims of a hoax or false cures) were removed from YouTube.[24] Meanwhile, Twitter reported removing 8,400 tweets and challenging 11.5 million accounts worldwide.[25] For Meta, by August 2021, more than 3,000 accounts, pages, and groups on Facebook had been removed for repeatedly violating rules about COVID-19 and vaccine misinformation. In fact, more than 20 million pieces of content have been removed from Facebook for COVID-19 misinformation, and more than 167 million pieces of debunked content relating to COVID-19 on Facebook were labeled and had their visibility reduced.

Section 230 is critical to promoting exactly this type of proactive content moderation activity on health misinformation. Numerous lawsuits have been filed against online services for removing or demoting the content this bill is intended to target. Some suits, such as Children's Health Defense's lawsuit against Facebook, argue that statements from politicians such as Senator Klobuchar which demand that companies conduct more content moderation are sufficiently coercive to online services that they convert private action into state action and therefore are restricted by the First Amendment.[26] Children's Health Defense's lawsuit is in front of the Ninth Circuit Court of Appeals right now. These types of suits also raise a litany of claims under theories

19 "Vaccine Misinformation Policy," YouTube Policies, accessed August 20, 2022, https://support.google.com/youtube/answer/11161123?hl=en&ref_topic=10833358.

20 "Updates to Our Work on COVID-19 Vaccine Misinformation," Twitter Safety, March 1, 2021, https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.

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

22 Stephanie Alice Baker, Matthew Wade, and Michael James Walsh, "The Challenges of Responding to Misinformation during a Pandemic: Content Moderation and the Limitations of the Concept of Harm," *Media International Australia* 177 (2020): 103–7, https://doi.org/10.1177%2F1329878X20951301.

23 YouTube Team, "Managing Harmful Vaccine Content on YouTube," September 29, 2021, https://blog.youtube/news-and-events/managing-harmful-vaccine-content-youtube/.

24 Neal Mohan, "Perspective: Tackling Misinformation on YouTube," YouTube, August 25, 2021, https://blog.youtube/inside-youtube/tackling-misinfo/.

25 "Updates to Our Work on COVID-19 Vaccine Misinformation," Twitter Safety, March 1, 2021, https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.

26 Children's Health Defense v. Facebook Inc., 20-cv-05787-SI (N.D. Cal. June 29, 2021). See also Doe v. Google LLC, 20-cv-07502-BLF (N.D. Cal. October 19, 2021); Huber v. Biden, 21-cv-06580-EMC (N.D. Cal. March 18, 2022); Loveland v. Facebook, 20-cv-6260-JMY (E.D. Pa. May 3, 2021); Atkinson v. Facebook Inc., 20-cv-05546-RS (N.D. Cal. December 7, 2020); Informed Consent Action Network ("ICAN") v. YouTube LLC, 20-CV-09456-JST, 2022 WL 278386, at *6 (N.D. Cal. January 31, 2022).

ranging from torts to unfair business practices. Section 230 immunity is a key defense for providers who remove this type of content.[27]

Hate Speech

On February 5, 2021, Senators Mark Warner (D-VA), Mazie Hirono (D-HI), and Amy Klobuchar (D-MN) introduced the Safeguarding against Fraud, Exploitation, Threats, Extremism and Consumer Harms (SAFE TECH) Act to "reform Section 230 and allow social media companies to be held accountable for enabling cyber-stalking, targeted harassment, and discrimination on their platforms."[28] In the press release announcing the bill's introduction, Senator Warner argued, "Section 230 has provided a 'Get Out of Jail Free' card to the largest platform companies even as their sites are used by scam artists, harassers and violent extremists to cause damage and injury."[29] In the same release, Jonathan A. Greenblatt, CEO of the Anti-Defamation League, stated, "Tech companies must be held accountable for their roles in facilitating genocide, extremist violence and egregious civil rights abuses. . . . The sweeping legal protections enjoyed by tech platforms cannot continue."[30]

According to the bill's sponsors, the bill would create a carve-out from Section 230 for "enforcement of stalking/cyberstalking or harassment and intimidation on the basis of protected classes." While the bill covers other activities, this essay will cover only the bill's impact as it relates to Section 230 and digital content that constitutes hate speech.

As with health misinformation, platforms have sought to moderate content promoting violence or hatred toward protected groups or individuals in various ways. In 2018, the European Commission noted that companies had removed over 70 percent of illegal hate speech on their platforms and were able to review almost 90 percent of content flagged as hate speech within 24 hours.[31] Mainstream platforms actively moderate hate speech content and have created internal regulatory infrastructure to address this issue, including developing and deploying algorithms that review and remove content, providing mechanisms for users to flag hateful content, and employing tens of thousands of moderators to review content.[32] All of these steps have led to an increase in the amount of hateful content found and flagged, particularly content flagged by the social media platforms themselves before it is reported by users.[33]

---

27 See, e.g., Doe v. Google LLC, 20-cv-07502-BLF (N.D. Cal. October 19, 2021) (regarding breach of contract); Daniels v. Alphabet Inc., 20-cv-04687-VKD (N.D. Cal. March 31, 2021) (regarding unjust enrichment and conversion); Loveland v. Facebook, 20-cv-6260-JMY (E.D. Pa. May 3, 2021) (regarding racketeer influenced and corrupt organizations, libel, promissory estoppel).

28 Mark R. Warner, "Warner, Hirono, Klobuchar Announce the SAFE TECH Act to Reform Section 230," press release, February 5, 2021, https://www.warner.senate.gov/public/index.cfm/2021/2/warner-hirono-klobuchar-announce-the-safe-tech-act-to-reform-section-230.

29 Warner, "Warner, Hirono, Klobuchar Announce the SAFE TECH Act."

30 Quoted in Warner, "Warner, Hirono, Klobuchar Announce the SAFE TECH Act."

31 Elizabeth Schulze, "EU Says Facebook, Google and Twitter Are Getting Faster at Removing Hate Speech Online," CNBC, February 4, 2019, https://www.cnbc.com/2019/02/04/facebook-google-and-twitter-are-getting-faster-at-removing-hate-speech-online-eu-finds--.html.

32 Richard Ashby Wilson and Molly K. Land, "Hate Speech on Social Media: Content Moderation in Context," *Connecticut Law Review* 52, no. 3 (February 2021): 1029–76, https://ssrn.com/abstract=3690616.

33 UNESCO and UN Office on Genocide Prevention and the Responsibility to Protect, "Addressing Hate Speech on Social Media: Contemporary Challenges," 2021, https://unesdoc.unesco.org/ark:/48223/pf0000379177.

On YouTube, in the period between July and December 2021, more than 200,000 videos with hate speech content were removed from the platform.[34] Meta took action on approximately 50 million pieces of hate speech content on its platforms between July and December 2021—40 million of which were on Facebook and 10 million on Instagram.[35] On Twitter, between January and July 2021, more than 6 million accounts were reported for hateful content.[36] The company took action on more than 1.1 million accounts, suspended more than 130,000 accounts, and removed more than 1.6 million pieces of content from the platform.[37]

However, providers' enforcement of their policies against hateful conduct and hate speech have been repeatedly challenged in court. While Twitter is not unique in this regard, it provides a good example of the type of content providers remove and the potential for user lawsuits as a result. In *Johnson v. Twitter, Inc.*, a user sued Twitter for enforcing its rules concerning violence and "wishing and hoping harm" after it took action against a tweet advocating that someone "take out Deray" (referring to civil rights activist DeRay McKesson).[38]

Similar suits were filed by other users who had tweets removed or accounts suspended, including *Jones v. Twitter, Inc.* (hate speech targeting Trevor Noah),[39] *Wilson v. Twitter, Inc* (hate speech targeting gay, lesbian, bisexual, and transgender people),[40] *Murphy v. Twitter, Inc*. (content misgendering and deadnaming transgender people),[41] *Verogna v. Twitter, Inc.* (hateful conduct and targeted harassment),[42] and *Martillo v. Twitter, Inc.* ("anti-zionist" content).[43] Section 230 has successfully protected online services' ability to enforce policies against hate speech and targeted harassment, including against claims of violating state tort laws and federal civil rights statutes.[44]

Child Sexual Abuse Material

On January 31, 2022, Senators Richard Blumenthal (D-CT) and Lindsey Graham (R-SC) introduced the Eliminating Abusive and Rampant Neglect of Interactive Technologies (EARN IT) Act, which the authors state "removes blanket immunity for violations of laws related to online child sexual abuse material (CSAM)."[45] The legislation had a multitude of bipartisan cosponsors at the time of its introduction. According to the bill's original authors, the bill creates a strong incentive for the tech industry to take online child sexual exploitation seriously by amending "Section 230 of

34 "Google Transparency Report: Featured Policies: Hate Speech," Google, accessed August 20, 2022, https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech.

35 "Hate Speech," Meta, accessed August 20, 2022, https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/.

36 "Rules Enforcement," Twitter, accessed August 20, 2022, https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jan-jun.

37 "Rules Enforcement," Twitter.

38 Cal. Super. No. 18CECG00078 (June 6, 2018).

39 Civil No. RDB-20-1963 (D. Md. October 23, 2020).

40 CIVIL ACTION No. 3:20-0054 (S.D.W. Va. June 16, 2020).

41 60 Cal.App.5th 12 (Cal. Ct. App. 2021).

42 2020 DNH 152 (D.N.H. 2020).

43 1:21-cv-11119-RGS (D. Mass. October 15, 2021).

44 See, e.g., Wilson v. Twitter, Inc., No. 3:20-cv-00495 (S.D.W. Va. September 17, 2020). This case dismissed Title II discrimination claims on the basis of plaintiff's "heterosexuality" and "Christianity" on grounds including Section 230.

45 Lindsey Graham, "Graham, Blumenthal Introduce EARN IT Act to Encourage Tech Industry to Take Online Child Sexual Exploitation Seriously," press release, January 31, 2022, https://www.lgraham.senate.gov/public/index.cfm/2022/1/graham-blumenthal-introduce-earn-it-act-to-encourage-tech-industry-to-take-online-child-sexual-exploitation-seriously.

the Communications Decency Act to remove blanket immunity from Federal civil, State criminal, and State civil child sexual abuse material laws entirely. Service providers will now be treated like everyone else when it comes to combating child sexual exploitation and eradicating CSAM, creating accountability."

CSAM is an urgent problem that requires technological, legislative, and social solutions. Some of the technological solutions include hash matching technology, automated detection software, and machine learning classifiers, which allow companies to identify and detect abuse content developed by technology companies. Tech companies also strive to eradicate child exploitation online and have taken a variety of approaches and developed a wide array of tools to combat CSAM on their platforms, including investing human and technical resources into building systems to detect, delete, and report CSAM content, responding to law enforcement requests, and supporting the National Center for Missing and Exploited Children (NCMEC) with financial and in-kind donations.

The industry has also taken joint action to fight CSAM through the creation of the Technology Coalition. The Technology Coalition was founded in 2006 as an industry response to the growth in CSAM online with a goal of promoting the development and adoption of tools to proactively detect CSAM, such as hash value scanning. The project has facilitated industry sharing of hash values for known images of CSAM for more than a decade, and currently has more than 25 providers. In 2020, Technology Coalition members provided 98 percent of all provider reports received by NCMEC.[46] This reporting is largely driven by deployment of automated detection tools, such as PhotoDNA, Content Safety API, and CSAI Match, that were developed and broadly shared by tech companies in furtherance of their goal to eradicate CSAM.

In addition, companies moderate their platforms by taking enforcement actions against accounts or content uploaded in contravention of CSAM policies. Social media platforms reported taking down materials of child sexual exploitation within the first hour after upload.[47] Between July and September 2021, Google disabled more than 140,000 accounts for CSAM violations, de-indexed almost 600,000 third-party web pages from Google Search, and reported to NCMEC more than 3 million pieces of content, including images, videos, URL links, and texts.[48] On its YouTube platform, the company removed 95,000 channels, more than 3 million videos, and almost 300 million comments for child safety reasons.[49] In total, more than 1.6 million CSAM hashes have been created and added to the NCMEC database to help detect previously identified CSAM.[50]

CSAM content was also extensively moderated by Meta and Twitter in 2021. Between July and December, more than 40 million pieces of content associated with child sexual exploitation and 3.6 million associated with child nudity and physical abuse were "actioned" on Facebook.[51] On

46 Tech Coalition, "The Technology Coalition Annual Report," accessed August 20, 2022, https://www.technologycoalition.org/annualreport/.

47 Alexandre de Streel et al., "Online Platforms' Moderation of Illegal Content Online," 2020, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf.

48 "Google Transparency Report: Google's Efforts to Combat Online Child Sexual Abuse Material," Google, accessed August 20, 2022, https://transparencyreport.google.com/child-sexual-abuse-material/reporting?hl=en.

49 "Google Transparency Report: Featured Policies: Child Safety," Google, accessed August 20, 2022, https://transparencyreport.google.com/youtube-policy/featured-policies/child-safety.

50 "Google Transparency Report: Featured Policies: Child Safety."

51 "Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation," Meta, accessed August 20, 2022, https://transparency.fb.com/data/community-standards-enforcement/child-nudity-and-sexual-exploitation/facebook/.

Instagram, more than 4 million pieces of content associated with child sexual exploitation and 1.5 million associated with child nudity and physical abuse were actioned.[52] In 2021, Twitter suspended more than 450,000 accounts and removed more than 6,000 pieces of content from its platform for child sexual exploitation reasons between January and June.[53]

While Senator Blumenthal and other EARN IT Act supporters say Section 230 prevents online services from taking steps to combat CSAM by shielding them from liability for user content, the level of industry response to this problem would be unthinkable without Section 230's protections. Today, providers use multifaceted approaches to detect CSAM, implementing advanced image- and video-matching technologies, new artificial-intelligence-based image classifiers, language and keyword algorithms and filters, URL scanning, and ways of processing third-party reports of CSAM content. Without Section 230, this type of proactive content moderation will increase the risk of liability for any CSAM content that is not caught or addressed quickly enough by providers' systems or personnel.

For example, while many of these detection methods are automated, they are possible only because at some stage in the process human content moderators review content to determine whether the content identified—for example, by a new AI-based tool—violates the provider rules and whether it triggers the federal requirement to report to NCMEC. Mistakes resulting in no action, such as judging a 16-year-old to be an 18-year-old, can give rise to levels of knowledge that may be actionable under federal and state CSAM laws. Conversely, the First Amendment may protect providers who take no proactive steps to identify CSAM.

Recent lawsuits brought against providers under the new sex trafficking exception to Section 230 provide a window into the risks of the return of the "moderator's dilemma." Plaintiffs bringing the suits argue that a "constructive knowledge" standard should apply to the analysis of whether a provider committed a violation of 18 U.S.C. § 1951, a prerequisite for the Section 230 exception to apply.[54] The question of whether this is the correct interpretation of the Fight Online Sex Trafficking Act (FOSTA) is currently in front of the Ninth Circuit in three separate cases.[55] If the Ninth Circuit adopts a constructive knowledge standard, providers are likely to alter the voluntary actions they currently take to address sex trafficking activity on their platforms. This response will likely include actions against CSAM, because CSAM can be evidence of sex trafficking activity. This situation is something that Congress should study before replicating FOSTA in other areas.

Disrupting the Narrative

Although popular political discourse stemming from the halls of Congress argues that big tech companies have turned a blind eye to the challenges presented by the prevalence of health disinformation, hate speech, and CSAM online, it is clear from public data, as well as publicly reported updates to content moderation policies, that the industry is making efforts to confront these challenges. Clearly, more work, research, and resources need to be invested by social media companies to ensure that dangerous content uploaded to online platforms is promptly moderated.

52 "Child Endangerment," Meta.

53 "Rules Enforcement," Twitter, accessed August 20, 2022, https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jan-jun.

54 See, e.g., *Doe v. Twitter.*

55 *Doe v. Twitter*, *Doe v. Reddit*, and *J.B. v. Craigslist*.

More importantly, Section 230's solution to the moderator's dilemma has spurred action rather than inaction and enabled companies take many of the efforts discussed above. Current proposals to amend Section 230 have the potential to upend freedom of speech principles and subordinate private content moderation to public law in a novel expansion of government power into the private sphere.[56] Federal policymakers would find more success in their efforts to tackle these significant policy concerns if they better understood the challenges and limitations of content moderation, the important role Section 230 plays in enabling moderation, and what policy solutions may exist beyond amendments to Section 230.

56 Kyle Langvardt, "Regulating Online Content Moderation," 2017, https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2018/07/Regulating-Online-Content-Moderation.pdf.