No. 21-1333

IN THE
Supreme Court of the United States

REYNALDO GONZALEZ, ET AL.,
Petitioners,

v.

GOOGLE LLC,
Respondent.

On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit

Brief of the Trust & Safety Foundation as *Amicus Curiae* in Support of Respondent

MARK W. BRENNAN
Counsel of Record
J. RYAN THOMPSON
GURTEJ SINGH
SOPHIE D. BAUM
HARSIMAR DHANOA
HOGAN LOVELLS US LLP
555 Thirteenth St., N.W.
Washington, DC 20004
(202) 637-5600
mark.brennan@hoganlovells.com

Counsel for Amicus Curiae

TABLE OF CONTENTS

TABLE OF AUTHORITIES

CASES:

STATUTES:

OTHER AUTHORITIES:

IN THE
Supreme Court of the United States

REYNALDO GONZALEZ, ET AL.,
Petitioners,

v.

GOOGLE LLC,
Respondent.

On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit

Brief of the Trust & Safety Foundation as *Amicus Curiae* in Support of Respondent

STATEMENT OF INTEREST[1]

The Trust & Safety Foundation ("TSF") is a 501(c)(3) nonprofit charitable organization that convenes key stakeholders to engage in interdisciplinary dialogue to help improve society's understanding and further the field of trust and safety ("T&S"). TSF does not make recommendations about how platform operators should run their businesses or develop their policies. TSF's sibling organization, the Trust & Safety Profes-

[1] No counsel for a party authored any part of this brief and no counsel or party made a monetary contribution intended to fund the preparation or submission of the brief. Only the amicus and its attorneys have paid for the filing and submission of this brief. Pursuant to Rule 37.3(a), all parties have granted blanket consent to the filing of amicus curiae briefs.

sional Association, is a 501(c)(6) nonpartisan membership association that supports the global community of professionals who develop and enforce principles and policies that define acceptable behavior and content online.

A thorough understanding of T&S operations and practices is essential for developing technology platforms, conducting research, fostering education, and crafting effective regulations and laws that benefit both users and platforms. This brief aims to explain the operational realities of content moderation under the current Section 230 regime and what is likely to occur if the scope of Section 230's protections is curtailed.

SUMMARY OF ARGUMENT

Content moderation is critical to the functioning of the modern internet. While the vast majority of users abide by content guidelines, the ability for user-generated content to be widely distributed on platforms also can invite instances of hate speech, harassment, and graphic violence. That is why virtually all platforms dedicate resources to limiting and removing objectionable content. The largest platforms' T&S teams comprise tens of thousands of workers and use sophisticated automated tools to maintain a consistent user experience.

In the United States, T&S teams can rely on Section 230 to help limit the amount of content that must be filtered or de-amplified. Even the most well-resourced T&S teams have neither the time nor resources to fully vet the legal risk of every piece of legally questionable content. Faced with the dilemma of leaving risk-creating content up or taking it down, T&S teams will be incentivized to take it down.

This burden cannot be borne by machines alone. To be sure, algorithmic tools can bear much of the content moderation workload. But they are not sophisticated enough to assess precisely the relative legal risk of all types of user-generated content. And of particular relevance here, limiting the scope of Section 230's protections for content recommendation tools could apply equally to content moderation tools because both tools select what user-generated content to display—and not display. Thus, limiting Section 230's protection for filtering software, like recommendation and content moderation algorithms, could be extremely burdensome to T&S teams and lead to even more content removal.

Limiting Section 230's protections could also harm user speech and platform competition. Platform design may shift from facilitating and encouraging user-generated content to limiting or foreclosing it. And smaller platforms may be unable to allocate the resources to manage increased intermediary liability risk, forcing them to leave the market, merge with a larger competitor, or not even enter it in the first place.

ARGUMENT

I. TRUST & SAFETY PLAYS KEY ROLES IN THE MODERN INTERNET.

A. The Trust & Safety Profession Has Grown Alongside the Internet and Is Critical to the Functioning of the Modern Internet.

Online services and technologies enable people to interact at an unprecedented scale. While these services and technologies drive innovation and growth and connect people in incredible ways, they can also serve as vehicles for harmful and unwanted behaviors

by some users. Online companies responded by investing in preventing, mitigating, minimizing, and—where possible and warranted—removing unwanted content or conduct.

T&S emerged as a field alongside the growth of the internet. For example, eBay was an early adopter of the term "trust and safety," where "trust" referred to trust among eBay users and of the company itself, and "safety" referred to keeping platform users safe. *See* Josh Boyd, *In Community We Trust: Online Security Communication at eBay*, 7 J. COMPUTER-MEDIATED COMMC'N (Apr. 2002). The term is now used to describe the teams that work to ensure that users are protected from harmful and unwanted experiences online or that may result offline due to online interactions.

T&S involves moderating user-generated content on websites or online platforms to ensure that such content complies with (1) all relevant laws and (2) the site's or platform's policies and guidelines. This may include, for example, developing and implementing content moderation policies and procedures to protect against online harassment, cyberbullying, and other forms of online abuse, as well as educating individuals on how to stay safe online. Along with post-hoc content moderation, many T&S teams also provide data to train machine-learning algorithms to block content preemptively and advise product teams on how to design abuse-resistant products and services.

T&S teams often share common key functions and roles. Content policy teams typically help develop content moderation policies and outline the content permitted on the platform. Platform operations teams handle day-to-day moderation activities and develop

scalable, efficient processes to support content moderation and policy enforcement.

Content moderation professionals are often (but not always) housed within broader operations teams. Harsha Bhatlapenumarthy, *Key Functions and Roles*, TR. & SAFETY PROF. ASS'N, https://bit.ly/3QICDdp. If a platform chooses to outsource moderation to an external vendor, the platform's operations team often manages the relationship with that vendor.[2] *Id.*

T&S legal teams respond to requests and issues as they arise; proactively identify and mitigate potential legal, reputational, and other risks; advise on compliance matters; and manage and respond to requests from law enforcement and regulators. They also establish and supervise processes for compliance with national law, such as the Digital Millennium Copyright Act for copyright removals in the United States and Germany's "NetzDG" law for hate speech and other categories of illegal content. *See* 17 U.S.C. § 512; Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [Netzwerkdurchsetzungsgesetz] [NetzDG] [Act to Improve Enforcement of the Law in Social Networks],

[2] Content design and strategy teams identify the optimal strategy for user-facing content and develop effective language to communicate with users. Data analytics teams build measurement methods to understand the extent of policy violations present on a platform. They analyze the effect of content moderation, proactive abuse detection, and various other efforts to curb violating content and behavior. Engineers develop machine learning models to scale and automate enforcement against policy-violating behavior. Compliance and law enforcement response teams review and carefully assess legal requests from law enforcement officials, ensuring compliance with applicable law and platforms' terms of service.

Sept. 1, 2017, BUNDESGESETZBLATT, TEIL I [BGBL I] (Ger.), translation at https://bit.ly/3IXUFGL.

The specific techniques and processes used for content moderation vary. But just as with other aspects of an online platform, T&S teams, policies, and tools must work at scale. The scale of content moderation can vary significantly depending on the size of the platform and the type of content it hosts.

Some platforms, such as social media websites with hundreds of millions or billions of users, require large T&S teams to review large amounts of potentially violative content in real-time. For example, Google reported in 2021 that its T&S team comprised nearly 22,000 employees. Google Submission in Response to Subcommittee Questions for the Record, *Hearing on Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation, Subcomm. on Commc'n & Tech. and the Subcomm. on Consumer Prot. & Com. of the H. Comm. on Energy & Com.*, 117th Cong. (2021), at 5, https://bit.ly/3He9sMf ("Supplemental Google Responses"). Facebook also reported in 2021 that it directly and indirectly employed 15,000 content moderators. Facebook Submission in Response to Subcommittee Questions for the Record, *Hearing on Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation, Subcomm. on Commc'n & Tech. and the Subcomm. on Consumer Prot. & Com. of the H. Comm. on Energy & Com.*, 117th Cong. (2021), at 39, https://bit.ly/3QU0R4w ("Supplemental Facebook Responses"). In the third quarter of 2022, Facebook's T&S team removed about 6.6 million pieces of content for violating its Bullying and Harassment policy, of

which about 2 million were removed after a user reported it. *Community Standards Enforcement Report: Bullying and Harassment*, META https://bit.ly/3IwCAz4 ("*Meta Bullying and Harassment Report*"). Roughly 1.5 million of those decisions were appealed. *Id.* In other words, while content moderation algorithms could automatically remove 4 million pieces of content, another 3.5 million human-based decisions had to be made to remove user-reported content and address appeals. By contrast, smaller platforms may rely on only a few employees or ask the community to enforce guidelines.

Content moderation helps fulfill a platform's mission and create a reliable user experience that can reflect user community preferences. *See, e.g.*, Supplemental Facebook Responses at 8 ("Billions of people use Facebook and Instagram because they have good experiences; they don't want to see hateful and violent content, our advertisers don't want to see it, and we don't want to see it."); Supplemental Google Responses at 30 ("Moderating content at scale is an immense challenge, but we see this as one of our core responsibilities[,] and we are focused on continuously working toward removing content that violates our policies before it is widely viewed."); Matt Perault, *Section 230 Reform: A Typology of Platform Power*, CPI ANTITRUST CHRON., at 18 (May 2021) ("Different approaches to content moderation enable users to make choices based on their moderation preferences.").

To support this mission, T&S teams emphasize consistency through policy design and enforcement. A policy that cannot be enforced consistently invites arbitrary decisionmaking, so T&S teams typically embed a quality assurance function to help ensure that

moderation decisions are made uniformly and fairly. Enforcement consistency improves trust among users, platforms, and their business partners. It also reduces confusion and allays fears of favoritism or unfair treatment.

While high consistency in policy enforcement can be achieved through careful planning and meticulous execution, even the best moderation systems are not perfect. As with any system in which policies must be created and enforced at scale, both errors and debatable judgment calls are inevitable. These issues arise for a few reasons. For one, content moderation very often relies on imperfect or incomplete information. T&S teams can only act on what they know or can infer about a piece of content, and the application of policy thus is not always a cut-and-dry case. For another, policies themselves are subjective judgments reflecting the values of a platform. Moreover, when millions of content moderation decisions are made daily, a 99.9% accuracy rate means that hundreds or thousands of errors are also made daily. And given the important role that online platforms play in society, judgment calls on whether or not to remove content will leave at least one set of constituents upset.

One obvious response to this conundrum is removing only indisputably illegal content. But, as is often the case in the T&S profession, the solution is not as simple as it first seems. T&S teams would need to evaluate each piece of content against an array of local, state, and federal laws. In other words, unsustainable business costs would result. It would also likely lead to an increase in unwanted—but legal—spam and abusive and harassing content. Moreover, that version of the internet would conflict with most users' preferences for at least some content moderation. *See,*

e.g., Shubham Atreja *et al.*, *What is the Will of the People? Moderation Preferences for Misinformation*, at 15 (Feb. 1, 2022), https://bit.ly/3jTauUo ("A majority of liberals preferred 'inform and reduce' actions on a few more articles than did a majority of conservatives. However, a majority of conservatives preferred removal of 22 articles while a majority of liberals preferred removal of only 15."); *Americans Support Free Speech Online but Want More Action to Curb Harmful Content*, KNIGHT FDN. (June 16, 2020), https://kng.ht/3kfb6nv ("Most people support the removal of false or misleading health information from social media. Amid the pandemic, 85% of Americans are in favor of this, and 81% support removing intentionally misleading information on elections or other political issues.").

T&S teams are constantly working to improve their content moderation practices and policies—all while balancing users' desire to express themselves with the needs of the broader community fostered by the platforms. This is a ceaseless endeavor that must constantly evolve with technology and society's changing preferences on the role of online speech.

B. All Content Moderation Approaches—from Highly Permissive to Restrictive—Rely on Section 230.

Online platforms' approaches to content moderation can vary significantly, reflecting each platform's goals and the community and culture they want to foster. For example, users' product reviews on Walmart.com must "[p]rovide specific details about . . . the product." *Ratings & Reviews Terms of Use*, WALMART, https://bit.ly/3WGNUNu. Scientific American prohib-

its users from "discuss[ing] topics unrelated to the article." *Scientific American Community Guidelines*, SCI. AM., https://bit.ly/3jISXyf.

Other platforms have more permissive policies or seek to take down only what the law requires. Social media platform Rumble, for example, describes itself as a pro-free-speech platform: "We are Rumble[.] . . . We create technologies that are immune to cancel culture." *About Us*, RUMBLE, https://bit.ly/3vyIXdB. Its content moderation policies essentially prohibit only illegal, pornographic, racist, violent, and terrorist content. *Website Terms and Conditions of Use and Agency Agreement*, RUMBLE, https://bit.ly/3IiNbgZ (last updated Oct. 27, 2022). Locals, an online community-building site that allows content creators to crowdfund and create communities, permits creators to independently police content in their "communities." Locals Community Guidelines list just a few categories of content that are prohibited on its platform: content that violates third-party rights (e.g., copyright), illegal content, "sexual activity," and "violence." *Community Guidelines*, LOCALS, https://bit.ly/3igiQVE (last updated Jan. 11, 2021).

The variety of content moderation approaches across online services is due in no small part to Section 230, which provides platforms with appropriately robust protection from liability for user-generated content. 47 U.S.C. § 230(c). It allows (but does not require) platforms to operate without putting a thumb on the scale of public discourse by censoring controversial or potentially unlawful (e.g., defamatory) content. In fact, for the platforms that provide significant latitude in permitted user content, the risk of intermediary liability would be greatest—but for Section 230.

But as detailed below in Section II, all platforms subject to U.S. law that host user-generated content will need to consider whether to—and may be forced to—more aggressively remove content if the Court curbs Section 230 protections.

C. Algorithmic Content Moderation Tools Are Valuable, Imperfect, and Reliant on Section 230.

The use of algorithms for content moderation has evolved significantly since the early days of the internet. At first, human moderators largely carried out content moderation, manually reviewing and removing inappropriate content. But as online platforms grew in size and scale, this became impractical, leading to increased use of automated systems.

These early algorithmic tools were relatively simple and focused on identifying and removing obviously unlawful or inappropriate (e.g., spam and hate speech) content. Over time, algorithms have become more sophisticated, incorporating machine learning techniques that allow them to adapt and improve over time.

Today, automated content moderation is essential for large social media platforms to manage the massive volume of user-generated content. Unwanted, harmful, and unlawful content can take many forms, such as spam, pornography, false or misleading information, and material that encourages violence or self-harm. Automated tools are now vital to filtering this content. For example, Meta reports that it relies on such tools—not user reports—to identify 90% of content that the platform removes for violating its hate speech policies. *Community Standards Enforcement Report: Hate Speech*, META, https://bit.ly/3jP2uUI.

While good and continuing to improve, the available tools are far from perfect for several reasons. For one, the performance of algorithms is intertwined with the human decisions that went into training them. Some of those decisions will be flawed due to the imperfect nature of human judgment—especially at the speed and scale of decisionmaking required by large online platforms. Those shortcomings are part of why human reviewers (and algorithm trainers) are still needed.

Algorithmic decisions are also probabilistic, and thus they will inevitably fail to perfectly identify all harmful posts as such (false negatives) and may also wrongly classify some benign posts as harmful (false positives). That is why algorithmic tools typically work best on content that is similar to content that humans have already reviewed in significant quantities. Machine-learning models require many examples (usually at least 10,000) for a model to accurately identify relevant content without a significant number of false positives and false negatives. *See generally Precision and Recall in Machine Learning*, JAVATPOINT, https://bit.ly/3Z5rcjC. So, for lower-volume or novel content, machine-learning models (1) often fail to work well or at all and (2) require time to deploy because many human-labeled examples must be made available to train the model.

Furthermore, machine-learning technology is still years away from being effective for reviewing all types of potentially violative content. For example, machine-learning models are limited in understanding context and nuance, especially for text-based content. This computational gap creates additional challenges for proactively detecting certain violation types. For example, as noted above, Meta automatically filters

90% of hate speech identified as violative, but that number drops to 68% for bullying and harassment. *Meta Bullying and Harassment Report*.

Due to these limitations, content moderation algorithms can, at best, identify materials for human evaluation, but they cannot consistently discern all content that is or could be illegal or contrary to a platform's policies. *See, e.g.*, Daphne Keller, *Amplification and Its Discontents: Why Regulating the Reach of Online Content is Hard*, 1 J. FREE SPEECH L. 227, 245 (2021) ("[N]either experience nor research suggests that algorithms can reliably distinguish legal from illegal content, outside of very limited cases."). For example, these algorithms cannot necessarily discern between a video of terrorist propaganda and a video of a journalist discussing a terrorist propaganda video.

Section 230 also protects from intermediary liability when websites use automated tools for content selection. *See, e.g.*, *Dyroff v. Ultimate Software Grp., Inc.*, 934 F.3d 1093 (9th Cir. 2019); *Force v. Facebook, Inc.*, 934 F.3d 53, 78-79 (2d Cir. 2019); *cf. Stratton Oakmont, Inc. v. Prodigy Servs. Co.*, No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995) (holding publisher liable of alleged libel, in part, because it controlled the content of its message boards through an "automatic software screening program"); S. Rep. No. 104-230, at 194 (1996) ("One of the specific purposes of [Section 230(c)] is to overrule *Stratton-Oakmont[, Inc.] v. Prodigy* . . . .").

This Court's decisions on the scope of Section 230 protection for recommendation algorithms could also affect algorithms used for content moderation and the functioning of T&S teams more broadly. The algo-

rithms at issue in this case involve filtering and picking content to provide the content for which the user is most likely looking. J.A.173. Content moderation algorithms similarly filter content disallowed on the platform. Functionally, there is no difference between amplifying a piece of content and de-amplifying all other content. Indeed, very often, the same technology that underlies content recommendation algorithms is used for content moderation algorithms.

Thus, if the Court adopts Petitioners' theory that YouTube's algorithmic recommendations do not receive Section 230 protection, any algorithmic content moderation action taken by a platform would appear to also fall within the purview of that holding. If YouTube can be held liable for video recommendations, then de-amplification of videos would seem to create liability too. This would place T&S teams in a bind.

II. LIMITING THE SCOPE OF SECTION 230 PROTECTION WOULD LEAD TO INCREASED RATES OF CONTENT REMOVAL, DESIGN CHANGES THAT HINDER SPEECH, AND REDUCED COMPETITION.

If the Court interprets Section 230 in a manner that limits online platforms' intermediary liability protections, T&S teams will need to alter their approaches to content moderation to mitigate the substantial increase in intermediary liability risk. The most likely risk mitigation technique would be to remove content that even comes close to creating liability for the platform because T&S teams do not have the time or resources to vet the legal risk created by individual pieces of user content. Content removal is a blunt response tool, but it has the virtue of being an action

that can be performed consistently, especially across larger T&S teams.

And because algorithmic content moderation tools are not sophisticated enough to identify such content with 100% certainty, T&S teams cannot comfortably offset the added burden with automated tools. Further, if the Court were to adopt Petitioner's theory that content recommendation algorithms do not receive Section 230 protection, such a decision could also effectively limit Section 230 protections for algorithmic content moderation tools.

This would dramatically hamper T&S teams' ability to combat various unlawful, unwanted, and harmful content. Platforms would be less likely to host controversial user speech. Some operators might redesign their platforms to limit significantly or even eliminate entirely user-generated content. And smaller platforms may be unable to afford the increased compliance and litigation costs, pushing them out of the market altogether.

A. Limiting Section 230's Protections Will Place Trust & Safety Teams in an Untenable Content Moderation Position.

Limiting the scope of Section 230's protections would force platforms to again face the so-called "moderator's dilemma"—the scenario that Section 230 was designed to eliminate.[3] Platforms would be forced to balance equally unattractive alternatives.

---

[3] *See Zeran v. Am. Online*, 129 F.3d 327, 330 (4th Cir. 1997) ("The amount of information communicated via interactive computer services is therefore staggering. The specter of tort liability in an area of such prolific speech would have an obvious chilling effect. It would be impossible for service providers to screen each of their millions of postings for possible problems. Faced with

One option would be for T&S teams to minimize the increased platform liability risk by scrutinizing every user post. But adopting this approach would effectively require pre-publication review of all content. Doing such reviews at scale would likely come with exorbitant costs and lengthy delays.

As a result, T&S teams would quickly be overwhelmed by assessing the accuracy and overall legality of every piece of user content hosted on their platforms. And even if such operational concerns could be overcome, this approach would require T&S teams to fully understand each piece of content's factual and social context to appropriately assess its legal risk. The largest platforms already struggle to identify, hire, and retain such experts at scale, leaving very few opportunities for smaller companies with fewer resources to compete for talent. Teams will likely respond to liability exposure by removing content for which they cannot interpret exact legal risk. This will inevitably result in systematic censorship of innocent people in a variety of languages and cultural groups.

Theoretically, content moderation algorithms could cut down on this burden. But practically, this approach would run into at least three problems. *First,* even for content that is fairly easy for algorithms to identify (e.g., some hate speech), these tools are imperfect. *See* supra Section I.C. They will be even less effective for more subtle types of violative content.

---

potential liability for each message republished by their services, interactive computer service providers might choose to severely restrict the number and type of messages posted. Congress considered the weight of the speech interests implicated and chose to immunize service providers to avoid any such restrictive effect.").

Second, for these algorithms to be effective, they would likely need to remove a large amount of non-violative, lawful content as well to ensure that the majority of violating content is captured, resulting in overbroad impingements on user speech. And *third*, if the Court concludes that content recommendation algorithms do not receive Section 230 protections, a likely corollary impact would be that decisions made by content moderation algorithms would also not be protected because they, too, filter content. *See supra* Section I.C. Platforms may, as a legal matter, be unable to rely on content moderation algorithms to mitigate the risk of user-generated content.

Another option would be for platforms to engage in no meaningful moderation, in which case they would not be liable for their users' content. Adopting this approach could remove the risk of liability for platforms in some cases, and it would certainly be more beneficial from a resource perspective.

But it would also transform the user experience on many major platforms. Indeed, the need for T&S teams arose from the rapid proliferation of illegal, offensive, and objectionable content that the internet enables. Many services have audiences that do not want this content and would not use the platform if it is not filtered.

To be sure, many T&S teams do remove content that raises potential liability for the platform while leaving a variety of questionable content available to users. But Section 230 currently provides flexibility to leave questionable content up so long as it is not clearly illegal. Eric Goldman and Jess Miers, *Online Account Terminations/Content Removals and the Benefits of*

Internet Services Enforcing Their House Rules, 1 J. FREE SPEECH L. 191 (2021) ("*House Rules*").

An alteration to Section 230's protections would force platforms to reconsider the balance of whether to risk moderating too little or too much. Current practice suggests that T&S teams are likely to shift toward more aggressive policies and enforcement to ensure consistency, removing any post that raises potential liability for the platform.

Platforms already take this approach toward certain content not protected by Section 230, such as intellectual property. 47 U.S.C. § 230(e)(2). This approach has led to a "notice and takedown" practice in which T&S teams review legal claims—often false or fraudulent ones—and attempt to decide which user speech should be silenced. This scheme is not only burdensome for reviewers but also ripe for abuse. Glyn Moody, *Copyright As Censorship: Abuse Of The DMCA To Try To Delete Online News Is Rampant*, TECHDIRT (May 24, 2022), https://bit.ly/3vxwC9F ("This [Digital Millennium Copyright Act] 'notice and takedown' system allows the copyright industry to . . . demand[] that [infringing content] should be taken down. . . . This unbalanced nature of the system makes it ripe for fraud, whereby people falsely claim to be the owner of copyright material in order to get it removed from a Web site. . . . An entire business sector, called 'reputation management,' has sprung up to offer this kind of service.").

As another example, Google reports that about half of the legal claims it has received under the EU's Right to Be Forgotten law are, effectively, attempts to suppress lawful online speech. False claimants have, Google reports, targeted some 2.3 million web pages

for removal. *Requests to delist content under European privacy law*, GOOGLE, https://bit.ly/3CjtgLe. While larger platforms may have the resources to review claims, smaller platforms often do not. Indeed, even under the much more limited removal obligations the U.S. provides under copyright law, some small companies report simply assuming that any legal accusation is valid—effectively creating a "heckler's veto" for anyone who contacts a platform to allege that a user's speech is illegal. Jennifer M. Urban *et al.*, *Notice and Takedown in Everyday Practice*, U.C. Berkeley Pub. L. Rsch. Paper No. 2755628 (2016).

In sum, if Section 230 protections are reduced, a likely outcome is for T&S teams to remove more aggressively any content that raises even the remote prospect of intermediary liability.

B. Platforms Are Likely to Remove Controversial and Legally Risky Speech More Often If Section 230 Protections Are Limited.

Online platforms are vital to disseminating information and opinions, even when they are controversial or go right up to the legal line—critical components for healthy debate and a well-informed citizenry. Section 230 allows platforms to host these controversial views without fear of legal reprisal. *See supra* Sections I.B, I.C. It also allows platform operators to create the digital community that they and their users prefer. Narrowing Section 230's protections would force platforms' T&S teams to take a far more aggressive approach toward content moderation to mitigate newly created liability vectors. *See supra* Section II.A. This will likely lead to the removal of

controversial speech vital to the functioning of civil society, significantly chilling free speech online.

Under Section 230, the risk of getting a content moderation decision wrong, in most cases, is a business issue for the platforms. Users become upset. Advertisers leave the platform. Bad publicity results. But these are, broadly speaking, not legal risks. Companies have wide latitude to determine what business and reputational risks they are willing to assume, and they can choose how to resource their T&S teams accordingly.

But without the robust intermediary liability protection that Section 230 currently provides, T&S teams will effectively be forced to remove content that *could* cross those lines well before they actually do. *See, e.g.*, *House Rules* at 204. The likely shift in content moderation approach would lead to broad curtailments of speech because T&S teams emphasize consistency in enforcement. *See supra* Section I.A. If teams are suddenly required to evaluate the individual legal risks of each piece of content, there's a tremendous incentive to default to removing entire categories of content where individual T&S team members (or their algorithmic tools) cannot make sufficiently consistent decisions quickly and at scale.

To limit a platform's legal exposure from a narrowed interpretation of Section 230, T&S teams would need to revise their platforms' content moderation policies to prohibit content that arguably comes close to the legal line or even prohibit discussion of controversial topics altogether. This approach would provide a buffer so that the platform's legal risk is minimized even if a moderator or algorithm makes the wrong decision (based on the platform's policies).

For example, assume a user post accuses a vaccine manufacturer of withholding key clinical trial data from regulators and the public. Currently, platform operators are unlikely to be held liable by the manufacturer for those claims. But if Section 230 protections are curtailed (e.g., by holding that Section 230 does not protect against algorithmic "boosting" of such content based on user interest), the platform's T&S team would (1) need to assess the factual accuracy of the claim and, (2) if the bases for such an allegation could not be substantiated quickly, remove the user's post to reduce the risk of liability to the platform.[4] In this scenario, the T&S team's decision to remove the user's *potentially violative* post may be a reasonable risk-mitigation decision.

Or, assume that a user published on a social media platform a particularly inflammatory and potentially defamatory post, which is then widely shared, leading to the platform's recommendation algorithms to list the post as a "trending" topic visible to a significant

---

[4] *See, e.g.*, *Zeran* at 333 ("If computer service providers were subject to distributor liability, they would face potential liability each time they receive notice of a potentially defamatory statement — from any party, concerning any message. Each notification would require a careful yet rapid investigation of the circumstances surrounding the posted information, a legal judgment concerning the information's defamatory character, and an on-the-spot editorial decision whether to risk liability by allowing the continued publication of that information. Although this might be feasible for the traditional print publisher, the sheer number of postings on interactive computer services would create an impossible burden in the Internet context. . . . Because service providers would be subject to liability only for the publication of information, and not for its removal, they would have a natural incentive simply to remove messages upon notification, whether the contents were defamatory or not.").

number of U.S. users. T&S teams currently would not necessarily need to worry about removing this recommended topic. But if Section 230 is interpreted as not protecting algorithmic recommendations, then the platform's T&S team would need to consider removing the post from trending topics. To limit liability risks, the T&S team may need to ban the user's account from ever "trending" on the platform, a practice that has recently received significant public criticism. *See, e.g.*, @bariweiss, TWITTER (Dec. 8, 2022, 7:20 PM), https://bit.ly/3Ck5oqB ("[T]eams of Twitter employees . . . prevent disfavored tweets from trending[] and actively limit the visibility of . . . trending topics . . . ."); Houston Keene, *Twitter Files: Unanswered questions remain after 'shadowban' revelations*, FOX NEWS (Dec. 9, 2022), https://fxn.ws/3QeBnyw.

While these are but two hypotheticals, it takes little to imagine the great variety of speech hosted by platforms that T&S teams would need to consider removing if intermediary liability protections are limited.

C. Platforms Will Be Less Likely to Host User Content If Section 230's Liability Protections Are Curtailed.

Litigation is expensive. And that holds true regardless of the outcome. *See Dombrowski v. Pfister*, 380 U.S. 479, 487 (1965) ("The chilling effect . . . [of litigation is] unaffected by the prospects of its success or failure."). So are compliance costs. Many operators will seek to avoid the increased costs and litigation risk by removing or forgoing user-generated content altogether and ceasing to offer trigger warnings. This will have a disproportionate effect on smaller platforms and reduce competition.

1. Platform Operators Will Design (or Redesign) Their Platforms to Limit Legal Liability from User Content.

Rather than focus on developing a product that consumers desire, with appropriate content moderation and community standards, many operators may default to designing their platforms within the constraints of T&S risk management and the avoidance of legal costs. For example, in jurisdictions in which intermediary liability protections like Section 230 are weak or unavailable, platform operators may opt to eliminate social media functionality altogether. *See, e.g.*, @mrwillhayward, TWITTER (Dec. 6, 2022, 3:09 AM), https://bit.ly/3Z1p8Jv ("A[n Australian] publisher is legally responsible for comments on tweets. So sometimes we have to turn them off if we're concerned a particular article is likely to draw a response that might be defamatory."). Not long ago, some companies took a similar approach, pulling out of Europe entirely, to avoid increased compliance costs and legal risks from the adoption of the General Data Protection Regulation. *See, e.g.*, Hannah Kichler, *US Small Businesses Drop EU Customers Over New Data Rule*, FIN. TIMES (May 23, 2018), https://on.ft.com/3G9V6uf.

Short of eliminating user engagement, platform operators may still limit functionality that some people want. For instance, platform operators often contextualize content by providing content notices, such as placing "age-appropriate" advisories on sensitive videos (e.g., putting a filter on a video to prevent playing until the user confirms their age). These tools help users and the parents of young users filter the content they view. If Section 230 protections are limited, op-

erators may be less likely to place these sorts of notices, to the detriment of consumers who may avoid a platform altogether if they cannot filter out undesirable content.

2. Limiting Section 230 Protections Will Reduce Competition Among Online Platforms.

Restricting Section 230 protections will have a disparate impact on smaller platforms. Section 230 is a pro-competitive law that places all platforms on relatively equal footing. *See* 47 U.S.C. § 230(b) ("It is the policy of the United States . . . to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation"). Altering the scope of Section 230's protections could balloon each platform's compliance costs, which smaller platforms would likely be unable to bear. And any expansion of the risk of expensive litigation could reduce competition, pushing smaller platforms to exit or forgo entering the market altogether. *See Race Tires Am., Inc. v. Hoosier Racing Tire Corp.*, 614 F.3d 57, 73 (3d Cir. 2010) ("[L]engthy and drawn-out litigation . . . may have a chilling effect on competitive market forces.").

User-generated content is growing at an astronomical rate. Under current protections, a startup might spend tens or hundreds of thousands of dollars on content moderation technology and staffing, while mid-sized platforms may spend millions. *See* Engine, *Startups, Content Moderation & Section 230* (2021), https://bit.ly/3i8dwE0. Given the volume of user-generated content hosted on even modest-sized platforms, newer and smaller internet services would likely face

far more litigation around their content moderation decisions than they could afford, especially as they scale up to compete with larger platforms. These significant, unrecoverable costs would impair the ability of new platforms to compete. Organizations such as local news companies, places of worship, and hobby and interest groups with websites with forums or user comments do not have the same resources to spend on content moderation as big corporations and will have less ability to take risks to compete.

There are a number of different approaches to content moderation that platforms select based on different business models and sizes of teams. *See* Robyn Caplan, *Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches*, DATA & SOC'Y, at 1 (Nov. 14, 2018), https://bit.ly/3WEaq9D. For instance, non-commercial sites, such as Wikipedia, often rely on community moderators and automated tools. *See, e.g.*, *Wikipedia: Automated moderation*, WIKIPEDIA, https://bit.ly/3jDMWmu. Modifying the protections of Section 230 could adversely affect these models in unexpected ways. These platforms may need to limit user content or require additional staff oversight. In many cases, platforms such as magazines, blogs, and newspapers will find it necessary to do away with their user-generated content (including comment sections) and instead provide only pre-approved content.

CONCLUSION

For all these reasons, the Court should affirm the lower court's decision.

Respectfully Submitted,

MARK W. BRENNAN
Counsel of Record
J. RYAN THOMPSON
GURTEJ SINGH
SOPHIE D. BAUM
HARSIMAR DHANOA
HOGAN LOVELLS US LLP
555 Thirteenth St., N.W.
Washington, DC 20004
(202) 637-5600
mark.brennan@hoganlovells.com

JANUARY 19, 2023 *Counsel for Amicus Curiae*