

NO. 23-30445

IN THE UNITED STATES COURT OF APPEALS
FOR THE FIFTH CIRCUIT

State of Missouri; State of Louisiana; Aaron Kheriaty; Martin Kulldorff; Jim Hoft;
Jayanta Bhattacharya; Jill Hines,

PLAINTIFFS-APPELLEES,

v.

Joseph R. Biden, Jr.; Vivek H. Murthy; Xavier Becerra; Department of Health &
Human Services; Anthony Fauci; et al.,

DEFENDANTS-APPELLANTS.

On Appeal from the United States District Court
for the Western District of Louisiana
Case No. 3:22-CV-1213
The Honorable Terry A. Doughty, District Court Judge

**BRIEF OF *AMICUS CURIAE* ELECTRONIC FRONTIER FOUNDATION
IN SUPPORT OF NEITHER PARTY**

David Greene
ELECTRONIC FRONTIER
FOUNDATION
815 Eddy Street
San Francisco, CA 94109
(415) 436-9333
davidg@eff.org

*Counsel for Amicus Curiae
Electronic Frontier Foundation*

SUPPLEMENTAL STATEMENT OF INTERESTED PARTIES

Pursuant to this Court’s Rule 29.2, the undersigned counsel of record for *amicus curiae* certifies that the following additional persons and entities have an interest in the outcome of this case:

1. Electronic Frontier Foundation, *amicus curiae*. *Amicus curiae* is a nonprofit organization recognized as tax exempt under Internal Revenue Code § 501(c)(3). It has no parent corporation and no publicly held corporation owns 10 percent or more of their stock.
2. David Greene, attorney for *amicus curiae*.

Dated: July 28, 2023

/s/ David Greene
David Greene

Counsel for Amicus Curiae
Electronic Frontier Foundation

TABLE OF CONTENTS

SUPPLEMENTAL STATEMENT OF INTERESTED PARTIES	ii
TABLE OF AUTHORITIES	v
STATEMENT OF INTEREST	1
INTRODUCTION	2
ARGUMENT	3
I. GOVERNMENT INVOLVEMENT IN CONTENT MODERATION INVOLVES COMPETING CONCERNS.....	3
A. Government Involvement in Content Moderation Raises Serious First Amendment Concerns.....	3
B. Government May Productively and Appropriately Contribute to Content Moderation Practices	4
II. FIRST AMENDMENT DOCTRINE REFLECTS THESE COMPETING INTERESTS AND REQUIRES COURTS TO DISTINGUISH BETWEEN APPROPRIATE AND INAPPROPRIATE GOVERNMENT COMMUNICATIONS	7
III. THIS COURT’S ANALYSIS MUST BE INFORMED BY A FULL UNDERSTANDING OF CONTENT MODERATION PROCESSES	13
A. Content Moderation is an Historic and Widely Employed Practice	14
B. Social Media Platforms Have Rules, Standards and Guidelines about What Content They Want and Don’t Want on Their Sites	16
C. Content Moderation Has Long Been and Remains a Fraught and Controversial Process	21

D. Social Media Platforms Commonly Consult With and Receive Advice and Feedback From Numerous External Sources, Including Governments25

CONCLUSION.....31

CERTIFICATE OF COMPLIANCE32

CERTIFICATE OF SERVICE33

TABLE OF AUTHORITIES

Cases

<i>Backpage.com v. Dart</i> , 807 F.3d 229 (7th Cir. 2015)	<i>passim</i>
<i>Bantam Books v. Sullivan</i> , 372 U.S. 58 (1963).....	<i>passim</i>
<i>Bose Corp. v Consumers Union</i> , 466 U.S. 485 (1984).....	3
<i>Carlin Comms., Inc. v. Mtn. States Tel. & Tel. Co.</i> , 827 F.2d 1291 (9th Cir. 1987)	12
<i>Hammerhead Enterprises, Inc. v. Brezenoff</i> , 707 F.2d 33 (2d Cir. 1983), <i>cert. denied</i> , 464 U.S. 892 (1983).....	12
<i>Kennedy v. Warren</i> , 66 F.4th 1199 (9th Cir. 2023)	11, 12, 13
<i>Miller v. California</i> , 413 U.S. 15 (1973).....	16
<i>Morse v. Frederick</i> , 551 U.S. 393 (2007).....	5
<i>Nat’l Rifle Ass’n of Am. v. Vullo</i> , 49 F.4th 700 (2d Cir. 2022)	11, 12
<i>O’Handley v. Weber</i> , 64 F.4th 1145 (9th Cir. 2023)	13
<i>Okwedy v. Molinari</i> , 333 F.3d 339 (2d Cir. 2003)	10
<i>Penthouse Int’l, Ltd. v. Meese</i> , 939 F.2d 1011 (D.C. Cir. 1991).....	10
<i>R.C. Maxwell Co. v. Borough of New Hope</i> , 735 F.2d 85 (3d Cir. 1984)	10, 11

Rattner v. Netburn,
 930 F.2d 204 (2d Cir. 1991) 10, 13

Rust v. Sullivan,
 500 U.S. 173 (1991)..... 5

Speech First, Inc. v. Cartwright,
 32 F.4th 1110 (11th Cir. 2022) 13

VDARE Found. v. City of Colorado Springs,
 11 F.4th 1151 (10th Cir. 2021), *cert. denied*, 142 S. Ct. 1208 (2022)..... 11, 12

Statutes

20 U.S.C. § 7114(d)(6)..... 5

22 U.S.C. § 4411(b) 5

Other Authorities

About the YouTube Priority Flagging Program, YouTube,
<https://support.google.com/youtube/answer/755433> 29

Adult Nudity and Sexual Activity, Facebook,
<https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/> 16

Article 19, *Sheikh Jarrah: Facebook and Twitter Silencing Protestors, Deleting Evidence* (May 10, 2021), <https://www.article19.org/resources/sheikh-jarrah-facebook-and-twitter-silencing-protests-deleting-evidence> 24

Cezary Podkul, *Scammers Are Using Fake Job Ads to Steal People’s Identities*, ProPublica (Oct. 26, 2021),
<https://www.propublica.org/article/scammers-are-using-fake-job-ads-to-steal-peoples-identities> 18

Child Safety Policy, YouTube,
<https://support.google.com/youtube/answer/2801999> 21

Community Guidelines, Instagram, <https://help.instagram.com/477434105621119> 16

Community Guidelines, Pinterest, <https://policy.pinterest.com/en/community-guidelines> 18, 19

Community Standards Enforcement Report, Meta,
<https://transparency.fb.com/data/community-standards-enforcement/>.....28

Danielle Blunt et al., *Posting Into The Void*, Hacking//Hustling (2020),
<https://hackinghustling.org/wp-content/uploads/2020/09/Posting-Into-the-Void.pdf>.....23

Derek E. Bambauer, *Against Jawboning*, 100 Minn. L. Rev. 51 (2015)13

Eric Goldman, *Content Moderation Remedies*, 28 Mich. Tech. L. Rev. 1 (2021)..... 15, 17

Facebook Treats Punk Rockers Like Crazy Conspiracy Theorists, Kicks Them Offline, EFF, <https://www.eff.org/takedowns/facebook-treats-punk-rockers-crazy-conspiracy-theorists-kicks-them-offline>.....24

Gettr – Terms of Use, Gettr (May 17, 2023), <https://gettr.com/terms>.....20

Hannah Bloch-Wehba, *Automation in Moderation*, 53 Cornell Int’l L.J. 41 (2020)..... 14

Hannah Denham, *Another Fake Video of Pelosi Goes Viral on Facebook*, Wash. Post (Aug. 3, 2020),
<https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/> 16

Jillian C. York & David Greene, *How to Put COVID-19 Content Moderation Into Context*, Brookings’ TechStream (May 21, 2020),
<https://www.brookings.edu/articles/how-to-put-covid-19-content-moderation-into-context/> 14

Jillian C. York, *Silicon Values: The Future of Free Speech Under Surveillance Capitalism* (Verso 2021).....22

Kari Paul, *Naked Protesters Condemn Nipple Censorship at Facebook Headquarters*, The Guardian (June 3, 2019)28

Kate Crawford & Tarleton Gillespie, *What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 New Media & Soc’y 410 (2014)27

Kate Klonick, *The New Governors: The People, Rules, And Processes Governing Online Speech*, 131 Harv. L. Rev. 1598 (2018)..... 15, 16, 17

Kevin Anderson, *YouTube Suspends Egyptian Blog Activist’s Account*,
 The Guardian (Nov. 28, 2007),
<https://www.theguardian.com/news/blog/2007/nov/28/youtubesuspendegyptianblog>22, 24

LinkedIn Professional Community Policies, LinkedIn,
<https://www.linkedin.com/legal/professional-community-policies> 19

Malachy Browne, *YouTube Removes Videos Showing Atrocities in Syria*,
 N.Y. Times (Aug. 22, 2017),
<https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html>; Anderson24

Megan Farokhmanesh, *YouTube Is Still Restricting and Demonetizing LGBT Videos—and Adding Anti-LGBT Ads to Some*, The Verge
 (June 4, 2018), <https://www.theverge.com/2018/6/4/17424472/youtube-lgbt-domentization-ads-algorithm>24

Meta Oversight Board, *UK Drill Music* (Jan. 2023),
<https://oversightboard.com/decision/IG-PT5WRTLW/>31

Meta, *How Stakeholder Engagement Helps Us Develop the Facebook Community Standards*, Meta (Jan. 18, 2023), <https://transparency.fb.com/en-gb/policies/improving/stakeholders-help-us-develop-community-standards/>26

Mike Masnick, *Content Moderation At Scale Is Impossible: Recent Examples Of Misunderstanding Context*, TechDirt (Feb. 26, 2021),
<https://www.techdirt.com/2021/02/26/content-moderation-scale-is-impossible-recent-examples-misunderstanding-context/>.....23

Misinformation Policies, YouTube Help,
<https://support.google.com/youtube/answer/10834785> 18

Moderator Code of Conduct, Reddit,
<https://www.redditinc.com/policies/moderator-code-of-conduct>26, 27

Muhammed Imran Shafique, *How to Spot and Avoid LinkedIn Job Scams*,
 LinkedIn (Nov. 26, 2022), <https://www.linkedin.com/pulse/how-spot-avoid-linkedin-job-scams-imran-shafique> 19

Naomi Appelman & Paddy Leerssen, *On ‘Trusted’ Flaggers*, 24 Yale J.L. & Tech. 452 (2022).....29

NHTSA, *Buzzed Driving is Drunk Driving*,
<https://www.nhtsa.gov/campaign/buzzed-driving>6

Policy: Terms of Use, Wikimedia,
https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use26

Promoting Hate Based on Identity or Vulnerability, Reddit,
<https://www.reddithelp.com/hc/en-us/articles/360045715951>20

Reddiquette, Reddit, <https://reddit.zendesk.com/hc/en-us/articles/205926439-Reddiquette>27

Reducing Hate And Disinformation Online, Change the Terms,
<https://www.changethetterms.org> 17

Report Content on Facebook, Facebook,
<https://www.facebook.com/help/181495968648557?rdrhc>.....27

Roblox Community Standards, Roblox, <https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards>20

Robyn Caplan, *Networked Governance*, 24 Yale J.L. & Tech. 541 (2022).....25

Role of Administrators and Moderators on Discord, Discord,
<https://discord.com/safety/360044103531-role-of-administrators-and-moderators-on-discord>26, 27

Rules Enforcement, Twitter (July 28, 2022),
<https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec>28

Ryan Mac, *Instagram Censored Posts About One of Islam’s Holiest Mosques, Drawing Employee Ire*, BuzzFeed News (May 12, 2021),
<https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque>.....23

Safety Partners, TikTok, <https://www.tiktok.com/safety/en-us/safety-partners/>26

Santa Clara Principles, <https://www.santaclaraprinciples.org/>4

Scams and Fraud Content, LinkedIn,
<https://www.linkedin.com/help/linkedin/answer/a1338803> 19

Sensitive Media Policy, Twitter (March 2023), <https://help.twitter.com/en/rules-and-policies/media-policy>..... 16

Seny Kamara et al., *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems*, Ctr. for Democracy & Tech. (2021), <https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems-updated-20220113.pdf> 15

Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. Pa. L. Rev. 11 (2006)..... 13

Taylor Wofford, *Twitter Was Flagging Tweets Including the Word “Queer” as Potentially “Offensive Content”*, Mic (June 22, 2017), <https://www.mic.com/articles/180601/twitter-was-flagging-tweets-including-the-word-queer-as-potentially-offensive-content> 25

The YouTube Team, *Our Ongoing Work to Tackle Hate*, YouTube (June 5, 2019), <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/> 18

Twitter Suspends Accounts of Palestinian Quds News Network, Al Jazeera (Nov. 2, 2019), <https://www.aljazeera.com/news/2019/11/2/twitter-suspends-accounts-of-palestinian-quds-news-network> 24

Website Terms and Conditions of Use and Agency Agreement, Rumble (Oct. 27, 2022), <https://rumble.com/s/terms> 21

YouTube Community Guidelines Enforcement, Google, <https://transparencyreport.google.com/youtube-policy/removals> 27, 29, 30

STATEMENT OF INTEREST¹

The Electronic Frontier Foundation (EFF) is a member-supported, nonprofit civil liberties organization that has worked for over 30 years to protect free speech, privacy, security, and innovation in the digital world. EFF, with over 35,000 members, represents the interests of technology users in court cases and broader policy debates surrounding the application of law to the Internet and other technologies. EFF frequently files briefs in cases addressing online intermediary content moderation, and studies and writes extensively on the issue.

¹ No party's counsel authored this brief in whole or in part. Neither any party nor any party's counsel contributed money that was intended to fund preparing or submitting this brief. No person other than *amicus*, its members, or its counsel contributed money that was intended to fund preparing or submitting this brief. Parties have consented to the filing this brief.

INTRODUCTION

Government co-option of the content moderation systems of social media companies is a serious threat to freedom of speech; but there are clearly times when it is permissible, appropriate, and even good public policy for government agencies and officials to noncoercively communicate with social media companies about the user speech they publish on their sites.

The applicable First Amendment test, from *Bantam Books v. Sullivan*, 372 U.S. 58, 71 (1963), recognizes these competing interests. The First Amendment forbids the government from intimidating or coercing a private entity to censor, whether the coercion is direct or subtle. *Id.* This has been an important principle in countering efforts to threaten and pressure intermediaries like bookstores, *see id.*, and credit card processors, *see Backpage.com v. Dart*, 807 F.3d 229, 230–31 (7th Cir. 2015), to limit others’ speech. But the test recognizes that not every communication to an intermediary about users’ speech is unconstitutionally coercive. *See Bantam Books*, 372 U.S. at 71-72.

The distinction between proper and improper speech is often obscure, leaving ample gray area for courts reviewing such cases to grapple in. But grapple in it they must.

The district court did not adequately distinguish between improper and proper communications in either its analysis or preliminary injunction. The

preliminary injunction is internally inconsistent with exceptions that seem to swallow many of its prohibitions. It does not provide adequate guidance to either the government or to anyone else seeking to hold the government to its proscriptions.

This Court must independently review the record and make the searching distinctions that the district court did not. *See Bose Corp. v Consumers Union*, 466 U.S. 485, 499-500 (1984) (holding that in First Amendment cases an appellate court must make an independent review of the whole record). This amicus brief aims to provide useful information about the competing interests involved and the environment of social media platform content moderation in which they must be applied.

ARGUMENT

I. GOVERNMENT INVOLVEMENT IN CONTENT MODERATION INVOLVES COMPETING CONCERNS

A. Government Involvement in Content Moderation Raises Serious First Amendment Concerns

Government involvement in private companies' content moderation processes raises serious human rights concerns. The recently revised Santa Clara Principles, of which *amicus curiae* is a co-author, specifically scrutinize "State Involvement in Content Moderation," and affirm that "state actors must not exploit or manipulate companies' content moderation systems to censor dissenters,

political opponents, social movements, or any person.”² “Special concerns are raised by demands and requests from state actors (including government bodies, regulatory authorities, law enforcement agencies and courts) for the removal of content or the suspension of accounts.”³

As will be discussed below, content moderation systems are fraught, and governmental manipulation of them to control public dialogue, silence disfavored voices, or blunt social movements raises classic First Amendment concerns. Governments have outsized influence to manipulate content moderation systems for their own political goals. This is especially problematic for government actors such as law enforcement whose communications may be inherently threatening or intimidating.

B. Government May Productively and Appropriately Contribute to Content Moderation Practices

Not every government communication to a social media platform is either improper or unwise. Indeed, for platforms seeking authoritative information, especially about government operations, the government may be its best source. Governments may be uniquely positioned to advise about the accuracy of user posts providing, for example, the location of polling places and polling hours,

² The Santa Clara Principles, <https://santaclaraprinciples.org> (last visited July 26, 2023).

³ *Id.*

natural disaster evacuation routes, and street closures.

Moreover, the government may have a role in publicly criticizing social media platforms when it disagrees with their content moderation actions, including when it believes that the platforms' policies or practices are endangering public safety.

As an uncontroversial general proposition, the federal government has a right to advance its viewpoints and attempt to convince others to adopt them. Indeed, it does this routinely. The National Endowment for Democracy, for example, was established for the United States to encourage other countries to adopt democratic principles. *See* 22 U.S.C. § 4411(b); *Rust v. Sullivan*, 500 U.S. 173, 194 (1991). And the Supreme Court not only approved of the former Safe and Drug-Free Schools and Communities Act of 1994, 20 U.S.C. § 7114(d)(6), which required schools receiving federal funds to have programs that “convey a clear and consistent message that ... the illegal use of drugs [is] wrong and harmful,” it empowered schools to censor on-campus, student speech that undermined that message. *See Morse v. Frederick*, 551 U.S. 393, 397 (2007).

The government frequently speaks to advance its initiatives. For example, the National Highway Transportation Authority currently has a campaign against “Buzzed Driving,” that is, driving after having consumed “even a small amount of alcohol,” when one’s blood alcohol content is below the legal limit of .08 and thus

noncriminal in most states.⁴ NHTSA is not disempowered from speaking out against or publicly admonishing those speakers who advocate for buzzed driving. NHTSA can criticize the speakers, call their information out as being false, and even encourage them to depublish what they had written; but the agency may not demand or coerce censorship. A social media platform that believed NHTSA's message was the better one and wanted to minimize posts on its site that encouraged "dangerous behavior" might choose to moderate the posts called out by NHTSA. A platform might also consider posts that asserted that buzzed driving is safe to be in violation of their misinformation policies.

Indeed, the district court itself seemed to find that in some situations the government had a duty to advocate its views to social media companies: it protested the government's failure to tell the large social media platforms that the Hunter Biden computer reporting was not false. ROA.26561. It also included exceptions to its order to highlight several categories of speech it apparently believed were appropriate for governmental intervention, namely "postings involving criminal activity or criminal conspiracies, national security threats, extortion, or other threats posted on its platform," "criminal efforts to suppress voting, to provide illegal campaign contributions, of cyber-attacks against election

⁴ NHTSA, *Buzzed Driving is Drunk Driving*, <https://www.nhtsa.gov/campaign/buzzed-driving> (last visited July 26, 2023).

infrastructure, or foreign attempts to influence elections,” “threats that threaten the public safety or security of the United States,” “promoting government policies or views on matters of public concern,” “postings intending to mislead voters about voting requirements and procedures,” and communications made “in an effort to detect, prevent, or mitigate malicious cyber activity.” ROA.26614-26615.

II. FIRST AMENDMENT DOCTRINE REFLECTS THESE COMPETING INTERESTS AND REQUIRES COURTS TO DISTINGUISH BETWEEN APPROPRIATE AND INAPPROPRIATE GOVERNMENT COMMUNICATIONS

Because of the competing concerns, courts have the difficult task of distinguishing between the government’s permissible communications—notifying, advising, convincing, mildly urging and encouraging—and the impermissible ones—coercing, threatening, compelling, intimidating, and other communications designed not merely to change the publisher’s mind but to cause them to censor even when in their own judgment they should and would not.

The proper test to assess whether the government has improperly effected censorship originates from *Bantam Books*, in which the Supreme Court found that the First Amendment prohibited not only direct censorship demands but also “system[s] of informal censorship” aimed at speech intermediaries. 372 U.S. at 71. The Supreme Court found that “the threat of invoking legal sanctions and other means of coercion, persuasion, and intimidation” against book distributors were enough to violate the book publishers’ First Amendment rights. *Id.* at 67.

In *Bantam Books*, a state commission engaged in such unconstitutional indirect coercion. The commission issued notices to book distributors that “certain designated books,” published by plaintiffs, were “objectionable for sale,” and that it was the commission’s “duty to recommend to the Attorney General prosecution of purveyors of obscenity.” *Id.* at 61-62. The commission also circulated the notices to local police, who visited the distributor “to learn what action he had taken.” *Id.* at 62-63. Predictably, the distributors stopped selling the books. *Id.* at 64. The Court found that the publishers had a First Amendment remedy against the state commission, even though it was the distributor’s action that directly harmed the publishers’ sales, and the government did not actually seize any books or prosecute anyone. *Id.* at 64 n.6.

Cases following *Bantam Books* have applied a totality of the circumstances analysis and identified numerous factors relevant to determining when government improperly pressured a speech intermediary to censor its users. This Court should view all these factors as relevant, with no single factor being essential, and weighted in favor of protecting First Amendment rights.

In *Backpage.com v. Dart*, the Seventh Circuit followed *Bantam Books* to distinguish between “attempts to convince and attempts to coerce,” the former being permissible and the latter forbidden. 807 F.3d 229, 230 (7th Cir. 2015). The case involved a sheriff’s campaign to shutdown Backpage.com’s adult section “by

demanding that firms such as Visa and MasterCard prohibit the use of their credit cards to purchase any ads on Backpage.” *Id.* On official letterhead, the sheriff demanded that the credit card companies “cease and desist” allowing payments for Backpage ads, citing the federal money-laundering statute. *Id.* at 231-32.

In deciding that the sheriff’s letter to the credit companies was coercion under *Bantam*, the court considered several factors.

First, the letter constituted an “implied threat” because it was written in the sheriff’s official capacity, “invoke[ed] the legal obligations of the financial institutions to cooperate with law enforcement,” and required further ongoing contact from the companies following the request. *Id.* at 236. The letter was not simply an effort to educate the recipients “about the nature and possible consequences of advertising for sex; he told them to desist or else.” *Id.* at 237. Because of the threatening language, the letter was not “a permissible attempt at mere persuasion.” *Id.* at 238.

Second, there was clear government *intent* to coerce Visa and Mastercard to cooperate. A strategic memo recommended appealing to the intermediaries’ interest in avoiding liability; the sheriff took credit in a press release for “compelling” the companies’ actions with his “demand”; and his office sent urgent communications to the companies following up on his letter which “imposed another layer of coercion due to [their] strong suggestion that the companies could

not simply ignore [the sheriff].” *Id.* at 232, 237.

Finally, though it was not a necessary finding,⁵ the court considered that the letter achieved its censorial goals. Visa and Mastercard had each received similar complaints from private citizens in the past, but severed ties with Backpage only two days after receiving the sheriff’s letter. *Id.* at 232-33. And even though Visa denied that it felt threatened, the court found an “obvious” causality between the sheriff’s letter and the credit card companies’ decisions to comply. *Id.* at 233.

Other courts have found numerous additional factors relevant.

In assessing whether a government communication was an implied threat, courts have considered whether it contained either a legal⁶ or economic⁷ threat. If legal, the court may inquire whether the government threatened criminal prosecution or merely a routine administrative action. For economic threats, courts will usually note where an intermediary in some way relies on the government’s

⁵ Indeed, the Seventh Circuit emphasized that “such a threat is actionable and thus can be enjoined even if it turns out to be empty—the victim ignores it, and the threatener folds his tent.” *Id.* at 231.

⁶ See *R.C. Maxwell Co. v. Borough of New Hope*, 735 F.2d 85, 88 (3d Cir. 1984) (finding no coercion where the government “brandish[ed] nothing more serious than civil or administrative proceedings under a zoning ordinance not yet drafted”); *Penthouse Int’l, Ltd. v. Meese*, 939 F.2d 1011, 1015 (D.C. Cir. 1991) (finding no coercion where letter contained “no threat to prosecute, nor intimation of intent to proscribe the distribution of the publications”).

⁷ See *Okwedy v. Molinari*, 333 F.3d 339, 343 (2d Cir. 2003); see also *Rattner v. Netburn*, 930 F.2d 204, 210 (2d Cir. 1991)).

favorable opinion for preference in securing future contracts, advantageous legislation, or some other public benefit.⁸

Courts also consider a communication’s “word choice and tone,” *Nat’l Rifle Ass’n of Am. v. Vullo*, 49 F.4th 700, 715 (2d Cir. 2022), as well as its context. *Kennedy v. Warren*, 66 F.4th 1199, 1209 (9th Cir. 2023) (“[A] full review requires us to analyze not only the tone of the letter but also the tenor of the overall interaction between Senator Warren and Amazon.”).⁹

Courts readily find coercion when the government agent has the authority to bring about the threatened consequence, and tend to find the communication permissible when the agent affirmatively genuinely disclaims any authority or intent to sanction.¹⁰

Courts are also more likely to find coercion when the government identifies

⁸ *R.C. Maxwell*, 735 F.2d at 87 (Citibank “denied that any explicit *quid pro quo* was contemplated or suggested, but admitted that Citibank’s desire to stay in the Council’s good graces was certainly a motivating factor.”).

⁹ *Compare NRA*, 49 F.4th at 717 (finding no coercion where guidance letters and press release were “written in an even-handed, nonthreatening tone and employed words intended to persuade rather than intimidate”), *with Backpage*, 807 F.3d at 231 (finding coercion where sheriff used “the legal term ‘cease and desist’”).

¹⁰ *See R.C. Maxwell*, 735 F.2d at 86 n.2 (“We are writing to solicit your personal assistance in order to alleviate an ongoing situation in our community by a professional agreement rather than legal procedures.”); *VDARE Found. v. City of Colorado Springs*, 11 F.4th 1151, 1157 (10th Cir. 2021), *cert. denied*, 142 S. Ct. 1208 (2022) (finding no coercion where the mayor issued public statement against hate speech and encouraged businesses to be attentive to the “types of events they accept” but acknowledged his lack of authority to restrict freedom of speech).

specific objectionable speech, and find no coercion when the government disapproves of a broad category of speech.¹¹

A government communication may also be found persuasive rather than coercive if instead of a threat it includes a reasoned argument for the censorship rather than a threat of penalty.¹²

Whether the government actor took further action or communication after the initial request is also relevant.¹³ Immediate and serious follow-up communications “continually reinforce[]” the request and thus cause the intermediary to reasonably interpret it as mandatory. *VDARE*, 11 F.4th at 1167.

Government requests that utilize an existing channel of communication to

¹¹ *Compare Carlin Comms., Inc. v. Mtn. States Tel. & Tel. Co.*, 827 F.2d 1291, 1295 (9th Cir. 1987) (finding coercion when deputy attorney advised telephone company to terminate the plaintiff’s “salacious” messaging service under threat of prosecution), *with NRA*, 49 F.4th at 716-17 (finding no coercion where government official sent guidance letters and issued press release encouraging regulated entities to “continue evaluating and managing their risks . . . that may arise from their dealings with the NRA or similar gun promotion organizations” in the wake of Parkland shooting (emphasis added)).

¹² *Hammerhead Enterprises, Inc. v. Brezenoff*, 707 F.2d 33, 36 n.2 (2d Cir. 1983), *cert. denied*, 464 U.S. 892 (1983) (finding no coercion where letter stated “Your cooperation in keeping this game off the shelves of your stores would be a genuine public service”); *Kennedy*, 66 F.4th at 1204 (finding no coercion letter stated that “[c]onspiracy theories about COVID-19 abound” and “have led to untold illnesses and deaths”).

¹³ *Kennedy*, 66 F.4th at 1209 (“An interaction will tend to be more threatening if the official refuses to take ‘no’ for an answer and pesters the recipient until it succumbs.”).

reach the intermediary, particularly where the intermediary welcomed or solicited the government’s input or expertise, tend to be seen as noncoercive.¹⁴

Lastly, courts are likely to view interactions with major power imbalances between the government actor and the targeted intermediary—such as where law enforcement is involved¹⁵—as more conducive to bullying or intimidation.¹⁶

III. THIS COURT’S ANALYSIS MUST BE INFORMED BY A FULL UNDERSTANDING OF CONTENT MODERATION PROCESSES

Because context is an important part of the First Amendment analysis, this court must understand how the often quixotic content moderation processes of social media platforms operate. At a bare minimum, this court must recognize that the notices, complaints, and take-down requests from the government are among

¹⁴ See *O’Handley v. Weber*, 64 F.4th 1145, 1163 (9th Cir. 2023) (finding no coercion where “OEC communicated with Twitter through the Partner Support Portal, which Twitter voluntarily created because it valued outside actors’ input”).

¹⁵ “[T]he emerging tactic of law enforcement officials in targeting ISPs with ‘requests’ that they take down websites that officials find problematic raises, in modern form, the threats to free expression implicit in any mechanism of prior restraint.” Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. Pa. L. Rev. 11, 76–77 (2006); see also *Kennedy*, 66 F.4th at 1210 (“A similar letter might be inherently coercive if sent by a prosecutor with the power to bring charges against the recipient, or if sent by some other law enforcement officer[.]”); *Rattner*, 930 F.2d at 210 (“unannounced visits by police personnel” are relevant to consideration of an implied threat).

¹⁶ *Speech First, Inc. v. Cartwright*, 32 F.4th 1110, 1123 (11th Cir. 2022) (“[T]he students targeted here are—for the most part—teenagers and young adults who, it stands to reason, are more likely to be cowed by subtle coercion than the relatively sophisticated business owners in those cases.”); see also Derek E. Bambauer, *Against Jawboning*, 100 Minn. L. Rev. 51, 103–106 (2015).

the millions of complaints, reports, and “flags” that social media companies routinely receive from a variety of stakeholders.

A. Content Moderation is an Historic and Widely Employed Practice

Social media platforms, at least from their point of mass adoption, have rarely published all legal speech submitted to their sites. Instead, they engage in content moderation: the use of policies, systems, and tools to decide what user-generated content or accounts to publish, remove, amplify, or manage.¹⁷ Large-scale, outsourced content moderation first emerged in the early 2000s.¹⁸

Platforms practice content moderation in phases: they define permissible and impermissible content; detect content that may violate their policies or the law; evaluate that content to determine whether it in fact violates their policies or the law; take an enforcement action against violative content; allow users to appeal or otherwise seek review of content moderation decisions that they believe are erroneous; and educate users about content moderation policies and their

¹⁷ See Hannah Bloch-Wehba, *Automation in Moderation*, 53 Cornell Int’l L.J. 41, 42, 48 (2020).

¹⁸ Jillian C. York & David Greene, *How to Put COVID-19 Content Moderation Into Context*, Brookings’ TechStream (May 21, 2020), <https://www.brookings.edu/articles/how-to-put-covid-19-content-moderation-into-context/>.

enforcement.¹⁹ In each phase, platforms make editorial judgments about what content they wish to allow or forbid on their services, or how to display or arrange it.

For example, during the definitional phase, some platforms develop a content policy, *i.e.*, a set of rules about what content is and is not allowed on their platforms.²⁰ Platforms may engage in significant internal discussion and debate, conduct internal and external research, and write multiple drafts before determining their content policies.²¹

Once a platform has decided, perhaps after deliberation and debate, during the evaluation phase that particular content violates its policies, the platform must decide what action to take in the enforcement phase. That is not a binary decision about whether to take down content or allow it to remain on a service, but also includes whether to change the manner or place in which content is displayed or to add the platform's own affirmative speech.²²

¹⁹ Seny Kamara et al., *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems*, Ctr. for Democracy & Tech. 9–11 (2021), <https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems-updated-20220113.pdf>.

²⁰ *Id.* at 9.

²¹ See Kate Klonick, *The New Governors: The People, Rules, And Processes Governing Online Speech*, 131 Harv. L. Rev. 1598, 1631-35 (2018).

²² See Eric Goldman, *Content Moderation Remedies*, 28 Mich. Tech. L. Rev. 1, 23–39 (2021) (describing various enforcement options).

B. Social Media Platforms Have Rules, Standards and Guidelines about What Content They Want and Don't Want on Their Sites

Social media platforms' content policies commonly prohibit users from posting speech that a platform believes is detrimental to its users and the public, its business interests, its editorial preferences, or all of these, even if that speech is legal. For example, many platforms ban legal, non-obscene sexual content, even though such speech enjoys First Amendment protection, *see Miller v. California*, 413 U.S. 15 (1973).²³

Content moderation differs from platform to platform.²⁴ Some platforms detect potentially violating content only after it is posted; others screen some or all content *ex ante*.²⁵ Platforms make different judgment calls about whether particular content violates their content policies, even if those policies are similar.²⁶ They use

²³ *See, e.g., Adult Nudity and Sexual Activity*, Facebook, <https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/> (last visited July 26, 2023).

²⁴ *Compare Community Guidelines*, Instagram, <https://help.instagram.com/477434105621119> (last visited July 26, 2023) (prohibiting nudity), *with Sensitive Media Policy*, Twitter (March 2023), <https://help.twitter.com/en/rules-and-policies/media-policy> (permitting “consensually produced adult nudity”).

²⁵ Klonick, *supra* n.21, at 1635.

²⁶ *See, e.g., Hannah Denham, Another Fake Video of Pelosi Goes Viral on Facebook*, Wash. Post (Aug. 3, 2020), <https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/> (reporting that TikTok, Twitter and YouTube removed a doctored video of Rep. Nancy Pelosi, while Facebook allowed it to remain with a label).

different methods to enforce their content policies, such as labeling content, placing interstitial warnings over it, or removing the ability to make money from it.²⁷ Some platforms allow users to appeal content moderation decisions, while others do not.²⁸

Many users prefer moderated platforms because they see benefits from moderation. Users may want to find or create affinity and niche communities dedicated to certain subject matters or viewpoints and exclude others. They may prefer environments that shield them from certain kinds of legal speech, including misinformation, pornography, hateful rhetoric and harassment, or simply speech that is off-topic or irrelevant.²⁹ And all users want services to filter out junk content or “spam.”

Users may want a service that only has highly trustworthy information and actively attempts to filter out misinformation. For example, one of the sites specified in the district court’s order, ROA.26612 n.2, Pinterest, a site designed to visually inspire creative projects, has “community guidelines” that “outline what

²⁷ Goldman, *supra* n.22, at 23–39.

²⁸ Klonick, *supra* n.21, at 1648.

²⁹ See, e.g., *Reducing Hate And Disinformation Online*, Change the Terms, <https://www.changetheterms.org> (last visited July 26, 2023) (campaign demanding improved content moderation against hate speech and disinformation).

we do and don't allow on Pinterest.”³⁰ Among the prohibited categories is “Misinformation.”³¹ Another site specified in the district court's order, YouTube, prohibits “misinformation” with serious risk of egregious harm.³² Moreover, its policy prohibiting promotion or glorification of Nazi ideology specifically prohibits misinformation in the form of Holocaust denial.³³

Content moderation also helps users avoid misinformation in the form of spam and scams. For instance, employment websites that allow employers to post job openings use spam and scam policies to combat, among other things, a growing trend of scammers using employment websites to steal applicants' identities in order to commit unemployment benefit fraud.³⁴ LinkedIn, for example, one of the platforms specified in the district court's order, ROA.26612 n.2, removes “phishing links, malware, known or suspected scam content, and fraudulent content and permanently restrict the accounts of known fraudsters or scammers,”

³⁰ *Community Guidelines*, Pinterest, <https://policy.pinterest.com/en/community-guidelines> (last visited July 26, 2023).

³¹ *Id.*

³² *Misinformation Policies*, YouTube Help, <https://support.google.com/youtube/answer/10834785> (last visited July 26, 2023).

³³ The YouTube Team, *Our Ongoing Work to Tackle Hate*, YouTube (June 5, 2019), <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/>.

³⁴ See Cezary Podkul, *Scammers Are Using Fake Job Ads to Steal People's Identities*, ProPublica (Oct. 26, 2021), <https://www.propublica.org/article/scammers-are-using-fake-job-ads-to-steal-peoples-identities>.

pursuant to its “Professional community policies.”³⁵ And LinkedIn encourages users who see scam postings to report them.³⁶

But misinformation is far from the only content category most platforms seek to exclude. Many platforms use content moderation to create environments that they believe are more user-friendly, including being focused on specific interests, prohibiting content that the platforms deem unsuitable for their purposes. Under these guidelines, Pinterest, for example, also reserves the right to remove several categories of speech: “Adult content,” “Exploitation,” “Hateful activities,” “Harassment and criticism,” “Private information,” “Self-injury and harmful behavior,” “Graphic Violence and Threats,” “Violent actors,” “Dangerous goods and activities,” “Harmful or Deceptive Products & Practices,” and “Impersonation.”³⁷ Pinterest also has special rules for comments users post on other users’ “Pins,” including a ban on “Irrelevant or non-purposeful material.”³⁸

³⁵ *LinkedIn Professional Community Policies*, LinkedIn, <https://www.linkedin.com/legal/professional-community-policies> (last visited July 26, 2023); *Scams and Fraud Content*, LinkedIn, <https://www.linkedin.com/help/linkedin/answer/a1338803> (last visited July 26, 2023).

³⁶ Muhammed Imran Shafique, *How to Spot and Avoid LinkedIn Job Scams*, LinkedIn (Nov. 26, 2022), <https://www.linkedin.com/pulse/how-spot-avoid-linkedin-job-scams-imran-shafique>.

³⁷ *Community Guidelines*, *supra* n.30.

³⁸ *Id.*

Many platforms also prohibit hate speech based on race, ethnicity, religion, and other characteristics. These kinds of rules are common, even on platforms that emphasize their commitment to free speech. For example, social media platform Gettr explains that it “holds freedom of speech as its core value and does not wish to censor your opinions,” while at the same time reserving the right to “address” content that attacks any religion or race.³⁹ Reddit’s content policy prohibits content that promotes “hate based on identity or vulnerability,” including race, religion, and national origin.⁴⁰

Many platforms prohibit content that praises or supports terrorism or other acts of violence. For example, Roblox, a rapidly growing online gaming platform, prohibits content that “incites, condones, supports, glorifies, or promotes any terrorist or extremist organization or individual.”⁴¹ Again, even platforms that tout their commitment to free speech often prohibit this content. For example, Rumble, a video sharing alternative to YouTube, bars content that “[p]romotes, supports or incites individuals and/or groups which engage in violence or unlawful acts, including but not limited to Antifa groups and persons affiliated with Antifa, the

³⁹ *Gettr – Terms of Use*, Gettr (May 17, 2023), <https://gettr.com/terms>.

⁴⁰ *Promoting Hate Based on Identity or Vulnerability*, Reddit, <https://www.reddithelp.com/hc/en-us/articles/360045715951> (last visited July 26, 2023).

⁴¹ *Roblox Community Standards*, Roblox, <https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards> (last visited July 26, 2023).

KKK and white supremacist groups and[/]or persons affiliated with these groups.”⁴²

Finally, some platforms also bar violent, sexual, or otherwise “inappropriate” content aimed at minors. YouTube’s child safety policy prohibits, among other things, “[f]amily friendly cartoons that target young minors and contain adult or age-inappropriate themes such as violence, sex, profanity, medical procedures, self harm, adult horror characters or other content intended to shock young audiences.”⁴³ It also prohibits “encouraging minors to do dangerous activities.”⁴⁴

C. Content Moderation Has Long Been and Remains a Fraught and Controversial Process

Content moderation controversies are neither a new problem nor limited to U.S. conservative politics.

In 2007, YouTube, only two years old at the time, shut down the account of Egyptian human rights activist Wael Abbas after receiving multiple reports that the account featured graphic videos of police brutality and torture.⁴⁵ YouTube’s

⁴² *Website Terms and Conditions of Use and Agency Agreement*, Rumble (Oct. 27, 2022), <https://rumble.com/s/terms>.

⁴³ *Child Safety Policy*, YouTube, <https://support.google.com/youtube/answer/2801999> (last visited July 26, 2023).

⁴⁴ *Id.*

⁴⁵ Kevin Anderson, *YouTube Suspends Egyptian Blog Activist’s Account*, The

community standards at the time stated that “[g]raphic, or gratuitous violence is not allowed.”⁴⁶ Just one year before, Abbas became the first blogger to receive the Knight International Journalism Award.⁴⁷

And government’s attempts to influence content moderation date back just as far: Abbas’s account was restored only after the U.S. State Department communicated with YouTube’s new owner, Google.⁴⁸

Content moderation remains a difficult and often fraught process that even the largest and best-resourced social media companies struggle with, often to the frustration of users. Even when using a set of precise rules or carefully articulated “community standards,” moderated platforms often struggle to draw workable lines between permitted and forbidden speech. Every online forum for user speech, not just the dominant social media platforms, struggles with this problem.

Platforms’ content moderation decisions are thus sometimes inconsistent or seemingly contrary to their own policies. Some of that is inevitable. Given the staggering amounts of content posted on platforms every day and the subjective

Guardian (Nov. 28, 2007), <https://www.theguardian.com/news/blog/2007/nov/28/youtubesuspendsegyptianblog>.

⁴⁶ *Id.*

⁴⁷ Jillian C. York, *Silicon Values: The Future of Free Speech Under Surveillance Capitalism* 25-27 (Verso 2021).

⁴⁸ *Id.*

judgment calls that some content moderation decisions require, platforms make mistakes in either moderating or failing to moderate content.⁴⁹

Beyond just mistakes, platforms have often aggressively removed content that is not prohibited by their content policies especially when attempting to minimize legal or reputational risks arising from government regulation or criticism. For example, many platforms responded to the enactment of the Allow States and Victims to Fight Online Sex Trafficking Act/Stop Enabling Sex Traffickers Act (“FOSTA”) by removing content by sex workers and sex worker advocates, even content that is not prohibited by FOSTA.⁵⁰ Government pressure to remove terrorist content from platforms has also led to over-removals of speech. For instance, in 2021, Instagram removed posts about one of Islam’s holiest mosques, Al-Aqsa, because its name is contained within the name of an organization the company had designated as a terrorist group.⁵¹

Platforms make a variety of contentious and seemingly erroneous decisions

⁴⁹ See Mike Masnick, *Content Moderation At Scale Is Impossible: Recent Examples Of Misunderstanding Context*, TechDirt (Feb. 26, 2021), <https://www.techdirt.com/2021/02/26/content-moderation-scale-is-impossible-recent-examples-misunderstanding-context/>.

⁵⁰ See Danielle Blunt et al., *Posting Into The Void*, Hacking//Hustling (2020), <https://hackinghustling.org/wp-content/uploads/2020/09/Posting-Into-the-Void.pdf>.

⁵¹ Ryan Mac, *Instagram Censored Posts About One of Islam’s Holiest Mosques, Drawing Employee Ire*, BuzzFeed News (May 12, 2021), <https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque>.

every day, and the impact of those decisions is felt across the political spectrum. In January 2021, Facebook’s updated policy to remove “harmful conspiracy theories” resulted in it disabling a punk rock band’s page because its name, Adrenochrome, is a chemical that was a central part of the QAnon conspiracy theory.⁵² YouTube has removed videos documenting atrocities in Syria and elsewhere under its graphic violence policy.⁵³ YouTube has also been accused of restricting and demonetizing LGBTQ+ content.⁵⁴ Twitter has been repeatedly criticized for moderating pro-Palestinian tweets, including removing those reporting on the events in Sheikh Jarrah in 2021⁵⁵ and blocking accounts associated with a major Palestinian news publication.⁵⁶ In 2017, users protested that Twitter had marked

⁵² *Facebook Treats Punk Rockers Like Crazy Conspiracy Theorists, Kicks Them Offline*, EFF, <https://www.eff.org/takedowns/facebook-treats-punk-rockers-crazy-conspiracy-theorists-kicks-them-offline> (last visited July 26, 2023).

⁵³ Malachy Browne, *YouTube Removes Videos Showing Atrocities in Syria*, N.Y. Times (Aug. 22, 2017), <https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html>; Anderson, *supra* n.45.

⁵⁴ Megan Farokhmanesh, *YouTube Is Still Restricting and Demonetizing LGBT Videos—and Adding Anti-LGBT Ads to Some*, The Verge (June 4, 2018), <https://www.theverge.com/2018/6/4/17424472/youtube-lgbt-domentization-ads-algorithm>.

⁵⁵ Article 19, *Sheikh Jarrah: Facebook and Twitter Silencing Protestors, Deleting Evidence* (May 10, 2021), <https://www.article19.org/resources/sheikh-jarrah-facebook-and-twitter-silencing-protests-deleting-evidence>.

⁵⁶ *Twitter Suspends Accounts of Palestinian Quds News Network*, Al Jazeera (Nov. 2, 2019), <https://www.aljazeera.com/news/2019/11/2/twitter-suspends-accounts-of-palestinian-quds-news-network>.

tweets containing the word “queer” as offensive.⁵⁷

D. Social Media Platforms Commonly Consult With and Receive Advice and Feedback From Numerous External Sources, Including Governments

To even hope for fairness and consistency in their decisions, social media companies need to draw on outside resources and expertise. This practice, which includes the use of trusted flagger programs, trust and safety councils, external stakeholder engagement teams, as well as as-needed consultations with individual and organizational experts, is widespread and often referred to as “networked governance.”⁵⁸ Governments are one component of this.

Meta, for example, says that it values stakeholder engagement to both develop and implement its community standards. With respect to misinformation, Meta says that “we engage extensively with experts and civil society stakeholders on topics such as state media, harmful health misinformation, and misinformation that may contribute to a risk of offline harm. Our team regularly speaks with academics and NGOs to provide visibility into how we develop and apply our

⁵⁷ Taylor Wofford, *Twitter Was Flagging Tweets Including the Word “Queer” as Potentially “Offensive Content”*, Mic (June 22, 2017), <https://www.mic.com/articles/180601/twitter-was-flagging-tweets-including-the-word-queer-as-potentially-offensive-content>.

⁵⁸ Robyn Caplan, *Networked Governance*, 24 Yale J.L. & Tech. 541, 542 (2022).

policies in these areas.”⁵⁹

TikTok has both “Safety Partners,” with which it “share[s] best practices, create programs, and exchange ideas on safety-related topics,” and “Advisory Councils” who “are made up of experts in youth safety, free expression, hate speech, and other safety topics. They work collaboratively with us to inform and strengthen our policies, product features, and safety processes.”⁶⁰

Platforms seek and receive feedback in a number of ways.

Some platforms structure their services so that some content moderation decisions are made by a broader community of users. Reddit and Discord rely on certain users to moderate content through the practice of “community moderation.”⁶¹ Reddit users manage and create thousands of communities, called “subreddits.” Although Reddit has an overriding content policy, a moderator makes the decisions within each community as guided by Reddit’s “Moderator

⁵⁹ Meta, *How Stakeholder Engagement Helps Us Develop the Facebook Community Standards*, Meta (Jan. 18, 2023), <https://transparency.fb.com/en-gb/policies/improving/stakeholders-help-us-develop-community-standards/>.

⁶⁰ *Safety Partners*, TikTok, <https://www.tiktok.com/safety/en-us/safety-partners/> (last visited July 26, 2023).

⁶¹ See, e.g., *Moderator Code of Conduct*, Reddit, <https://www.redditinc.com/policies/moderator-code-of-conduct> (effective Sept. 8, 2022); *Role of Administrators and Moderators on Discord*, Discord, <https://discord.com/safety/360044103531-role-of-administrators-and-moderators-on-discord> (last visited July 26, 2023); *Policy: Terms of Use*, Wikimedia, https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use (last visited July 26, 2023).

Code of Conduct.”⁶² Discord employs a similar model.⁶³ Each site thereby empowers some users to remove and down-rank other users’ speech if that speech is against the community’s rules.⁶⁴ Platforms also commonly rely on users to report content that violates the law or the platform’s policies.⁶⁵

Most platforms, even those which do not employ community moderation, allow users to report or “flag” content they believe violates the platforms’ rules or standards.⁶⁶ Indeed, such flags may account for a large amount of content moderation decisions. In the first quarter of 2023, YouTube removed 6.5 million videos, 360,000 of which were flagged by users.⁶⁷ During that same period, 34.8% of the 6.9 million posts Facebook actioned for bullying and harassment was reported by users, while 18% of the 10.7 million posts they actioned for hate

⁶² Reddit, *supra* n.61.

⁶³ Discord, *supra* n.61.

⁶⁴ See, e.g., *Reddiquette*, Reddit, <https://reddit.zendesk.com/hc/en-us/articles/205926439-Reddiquette> (last visited July 26, 2023).

⁶⁵ See, e.g., *Report Content on Facebook*, Facebook, <https://www.facebook.com/help/181495968648557?rdrhc> (last visited July 26, 2023).

⁶⁶ See generally Kate Crawford & Tarleton Gillespie, *What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 *New Media & Soc’y* 410 (2014).

⁶⁷ *YouTube Community Guidelines Enforcement*, Google, <https://transparencyreport.google.com/youtube-policy/removals> (last visited July 26, 2023).

speech was reported by users.⁶⁸ In the second half of 2021, users reported over 11 million accounts to Twitter as having violated at least one of its rules.⁶⁹

Platforms also commonly seek and/or receive input from civil society groups, activists, and other stakeholders who are not necessarily their users. This input may urge platforms to take content down *or* to put it back up. For example, *amicus* joined efforts pressuring Facebook to end its ban on pictures of female nipples.⁷⁰

Platforms may themselves reach out to a government agency or official when they perceive them as being authoritative. For example, the district court found it significant that Facebook solicited the opinion of the Centers for Disease Control regarding the accuracy of posts promoting ivermectin in treating COVID. ROA.26510–26511.⁷¹

⁶⁸ *Community Standards Enforcement Report*, Meta, <https://transparency.fb.com/data/community-standards-enforcement/> (last visited July 26, 2023).

⁶⁹ *Rules Enforcement*, Twitter (July 28, 2022), <https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec>.

⁷⁰ Kari Paul, *Naked Protesters Condemn Nipple Censorship at Facebook Headquarters*, *The Guardian* (June 3, 2019), <https://www.theguardian.com/technology/2019/jun/03/facebook-nude-nipple-protest-wethenipple>.

⁷¹ This appears to have been on Facebook's initiative and not in response to a CDC demand. There is no indication in the district court's ruling that the CDC first contacted Facebook about such posts.

Many platforms have potentially problematic “trusted flagger” programs in which certain groups and individuals enjoy “some degree of priority in the processing of notices, as well as access to special interfaces or points of contact to submit their flags.” Naomi Appelman & Paddy Leerssen, *On ‘Trusted’ Flaggers*, 24 Yale J.L. & Tech. 452, 453 (2022).

YouTube’s “Priority Flagger” program prioritizes complaints from certain entities, including individuals and NGOs, that are “particularly effective” at notifying YouTube of violative content.⁷² From January to March of this year, the platform removed 43,056 videos due to these complaints.⁷³

The government is thus among many entities and individuals that notify, contact, urge, or encourage the various social media companies to moderate user posts. And it is also among the entities and individuals to which the platforms will themselves reach out when they are seeking expertise. YouTube, for example, includes government agencies among those who may be “priority flaggers”;⁷⁴ in the first quarter of this year, seven videos were removed as a result of government flags, as compared to over 400,000 takedowns originating from non-governmental

⁷² *About the YouTube Priority Flagger Program*, YouTube, <https://support.google.com/youtube/answer/7554338> (last visited July 26, 2023).

⁷³ Google, *supra* n.67.

⁷⁴ Youtube, *supra* n.72.

external flags.⁷⁵

Of course, governmental participation in content moderation processes raises First Amendment issues not present with non-governmental inputs. As the *Bantam Books* test described above makes clear, government entities are rightfully significantly limited in how they can influence publication decisions.

As just one recent example of the hazards of establishing flagging pathways between governments and platforms, the Meta Oversight Board reviewed the decision by Meta to remove a musical track from Instagram, after Instagram received a request from London Metropolitan Police to review all content containing the track. The police opined that the song contained a veiled threat of gang violence and could lead to further violence. Ultimately, 165 pieces of content containing the track were removed. Through freedom of information laws, the Oversight Board learned that the police had made 286 requests to various social media companies to remove tracks of the same genre, drill music, resulting in 255 removals across the platforms. The Oversight Board overturned the decision, and also recommended that Meta increase transparency around government take-down requests and regularly review data on content moderation decisions prompted by

⁷⁵ Google, *supra* n.67.

government requests for systemic bias.⁷⁶

CONCLUSION

This court must independently examine the record and carefully distinguish the proper from improper contacts alleged in this case, with an understanding of the context in which social media companies ultimately decide whether and how to take action on any of the millions users posts they publish every day.

Dated: July 28, 2023

By: /s/ David Greene

David Greene

Counsel of Record

ELECTRONIC FRONTIER
FOUNDATION

815 Eddy Street

San Francisco, CA 94109

(415) 436-9333

davidg@eff.org

Counsel for Amicus Curiae

Electronic Frontier Foundation

⁷⁶ Meta Oversight Board, *UK Drill Music* (Jan. 2023), <https://oversightboard.com/decision/IG-PT5WRTLW/>.

CERTIFICATE OF COMPLIANCE

1. This brief complies with the type-volume limitation of Fed. R. App. P. 29(a)(5) because:

this brief contains 6,428 words, excluding the parts of the brief exempted by Fed. R. App. P. 32(a)(7)(f), or

this brief uses a monospaced typeface and contains [less than 650] lines of text, excluding the parts of the brief exempted by Fed. R. App. P. 32(a)(7)(f)

2. This brief complies with the typeface requirements of Fed. R. App. P. 32(a)(5), and the type style requirements of Fed. R. App. P. 32(a)(6) because:

this brief has been prepared in a proportionally spaced typeface using Microsoft Word for Mac in 14 point Times New Roman font, or

this brief has been prepared in a monospaced typeface using [name and version of word processing program] with [number of characters per inch and name of type style].

Dated: July 28, 2023

By: /s/ David Greene
David Greene

Counsel for Amicus Curiae
Electronic Frontier Foundation

CERTIFICATE OF SERVICE

I hereby certify that the foregoing was filed electronically with the Clerk of the Court for the United States Court of Appeals for the Fifth Circuit by on July 28, 2023 and will be served electronically upon all counsel.

Dated: July 28, 2023

By: /s/ David Greene
David Greene

Counsel for Amicus Curiae
Electronic Frontier Foundation