No. 21-1496

In the Supreme Court of the United States

TWITTER, INC., PETITIONER,

v.

MEHIER TAAMNEH, ET AL., RESPONDENTS.

ON WRIT OF CERTIORARI
TO THE UNITED STATES COURT OF APPEALS
FOR THE NINTH CIRCUIT

BRIEF OF COMPUTER & COMMUNICATIONS INDUSTRY ASSOCIATION, NETCHOICE, SOFTWARE & INFORMATION INDUSTRY ASSOCIATION, DEVELOPERS ALLIANCE, CHAMBER OF PROGRESS, INTERNET INFRASTRUCTURE COALITION, CONSUMER TECHNOLOGY ASSOCIATION, AND ACT | THE APP ASSOCIATION AS AMICI CURIAE IN SUPPORT OF PETITIONER

Scott A. Keller
Counsel of Record
Steven P. Lehotsky
Jeremy Evan Maltz
LEHOTSKY KELLER LLP
200 Massachusetts Ave., NW
Washington, DC 20001
(512) 693-8350
scott@lehotskykeller.com

Counsel for Amici Curiae

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page(s)

Case

Statute

Other Authorities

INTEREST OF AMICI CURIAE

The Computer & Communications Industry Association (CCIA) is an international, not-for-profit trade association representing a broad cross section of communications, technology, and Internet industry firms that collectively employ more than 1.6 million workers, invest more than $100 billion in research and development, and contribute trillions of dollars in productivity to the global economy.[1] For 50 years, CCIA has promoted open markets, open systems, and open networks. CCIA believes that open, competitive markets and original, independent, and free speech foster innovation.

NetChoice is a national trade association of e-commerce and online businesses that share the goal of promoting convenience, choice, and commerce on the Internet. For over a decade, NetChoice has worked to increase consumer access and options via the Internet, while minimizing burdens on small businesses that are making the Internet more accessible and useful.

The Software & Information Industry Association (SIIA) is the principal trade association for the software and digital information industries. SIIA's membership includes more than 400 software companies, search engine providers, data and analytics firms, and digital publishers that serve nearly every segment of society, including business, education, government, healthcare, and consumers. It is dedicated to creating a healthy environment for the

[1] Petitioner and Respondents have consented to the filing of this brief. In accordance with Rule 37.6, no counsel for any party has authored this brief in whole or in part, and no person or entity, other than amici or their counsel, has made a monetary contribution to the preparation or submission of this brief.

creation, dissemination, and productive use of information.

The Developers Alliance is a non-profit corporation that advocates for software developers. Its corporate mission is to "[a]dvocate on behalf of developers and the companies that depend on them, support the industry's continued growth, and promote innovation."[2] Alliance members include industry leaders in consumer, enterprise, industrial, and emerging software, and a global network of more than 75,000 developers.[3]

Chamber of Progress is a non-profit tech industry coalition devoted to a progressive society, economy, workforce, and consumer climate. It backs public policies that will build a fairer, more inclusive country in which all people benefit from technological leaps.

The Internet Infrastructure Coalition (i2Coalition) is an international, not-for-profit trade association founded in 2012 that supports and represents the companies that build and operate the infrastructure of the Internet. The broad and diverse membership of the i2Coalition includes cloud providers, data centers, web hosting companies, domain registries and domain registrars, content delivery networks, Internet exchange point providers, and network protection services.

As North America's largest technology trade association, CTA® is the tech sector. Their members are the world's leading innovators—from start-ups to global brands—helping support more than 18 million American

---

[2] About Us, Developers Alliance, https://bit.ly/3gJhqT3.

[3] A list of Developers Alliance members is available at https://bit.ly/3GSAiJW.

jobs. CTA owns and produces CES®—the most influential tech event in the world.

ACT | The App Association is a global trade association for small- and medium-sized technology companies. Its members are entrepreneurs, innovators, and independent developers within the global app ecosystem that engage with verticals across every industry. The App Association works with and for its members to promote a policy environment that rewards and inspires innovation while providing resources that help them raise capital, create jobs, and continue to build incredible technology.

Amici have witnessed firsthand how all websites and applications that publish, disseminate, or transmit user-generated speech online (what this brief will call "online services") have responded to the dangerous and objectionable content that makes its way to their websites and applications. As the amount of content on the Internet has grown exponentially, online services have reacted by adopting increasingly advanced means of removing and reducing the spread of content that violates their policies (what this brief will call "content moderation").

SUMMARY OF ARGUMENT

I. Content moderation is a necessary, complex, and unending task that operators of online services willingly undertake to maintain functioning online communities. Every day, people around the world create quintillions of bytes of data on the Internet. The Internet therefore comprises a wide array of online services, each fostering its own unique community of users through editorial policies. From single-issue web forums to social media websites disseminating various forms of speech, online services implement distinct editorial policies. These editorial policies determine what expression an online service deems acceptable to publish and disseminate.

Though these online services vary in what kinds of communities they seek to foster, they share a common goal: their users must consider the online services beneficial places to find content and to share content with others. At the most fundamental level, that means users should be comfortable with the kinds of content they will encounter on these websites and applications.

To make their communities useful to users and to comply with the law, online services remove or reduce the spread of content that their communities deem objectionable. In other words, online services engage in "content moderation." Sometimes, what content an online service deems objectionable depends on the nature of the community itself. Communities dedicated to particular hobbies, for instance, may prohibit off-topic discussions of other hobbies or expression that denigrates the hobby. At the same time, there is content widely recognized to be objectionable (or even illegal) across online communities: spam, scams, pornography, material harmful to children,

harassment, heckler's vetoes, child sexual abuse material, and copyright infringement, just to name a few.

The market has repeatedly acknowledged that online services' content moderation is necessary, as advertisers and users have boycotted online services that insufficiently moderate objectionable content. In the last four years, prominent online services have faced organized advertiser boycotts in protest of online services' perceived inadequate removal of harmful content like hate speech. Just as users do not want to encounter certain content, advertisers do not want their advertisements placed next to certain content. Each boycott caused these online services to lose millions of dollars, prompting further investment into more sophisticated approaches to content moderation. Today, many advertisers are publicly reallocating their advertising budgets based on what they perceive to be risks to their brands arising from changes in Twitter's approach to content moderation.

II. Content moderation is extraordinarily difficult—and costly. Online services constantly refine their content-moderation efforts to address new problems.

The Internet contains countless pieces of expressive content, and some portion of that content will inevitably fail to comply with online services' policies. Although the amounts of harmful and objectionable content are small percentages of all content online, they are large in absolute terms. For example, over a six-month period in 2020, seven online services removed nearly six billion pieces of harmful content (out of countless pieces of content overall), including spam, hate speech, harassment, material

harmful to children, and graphic violence including terrorist speech.[4]

As the Internet has grown and evolved, so too have online services. The relatively simple chat rooms and web forums of the 1990s have been joined by modern social media websites disseminating, displaying, and arranging billions of pieces of expressive content. At the same time, technological advances have provided more opportunities for objectionable content.

Online services try to remove harmful and objectionable content as quickly as possible—ideally before many (if any) people see it. They rapidly determine whether submitted content is so objectionable that it should be removed under the online services' policies or whether there is a way to effectively reduce the content's spread.

Succeeding at content moderation therefore requires a continuous, iterative process of improvement. The global Internet's sheer scale and the complexities of human interaction (in all languages and across many cultures) complicate online services' efforts. Online services have become able to quickly and effectively parse individual pieces of content's context to determine whether and how to continue disseminating that content. Online services may determine that not all potentially harmful content warrants removal and, conversely, not all policy-compliant content is worthy of prominent presentation. And there will not always be "correct" answers about how to address specific content.

Compounding online services' obligations, every advancement in speech dissemination and content

---

[4] NetChoice, By the Numbers: What Content Social Media Removes and Why 2-3, https://bit.ly/3Gn54Hj [hereinafter "By the Numbers"].

moderation prompts an escalatory response from those seeking to evade online services' content moderation. For example, social media websites often use computer algorithmic sorting and presentation of content, and they offer users the ability to easily upload (and livestream) videos and images, at massive scale. These features allow modern online services to foster their unique communities, yet they also allow a small percentage of users to push harmful content on others.

Despite content moderation's inherent difficulty, online services' efforts have been extremely successful. Among the online services that report such statistics, they remove 90% or more of harmful content before many users see it. For pro-terrorist speech specifically, online services have worked collaboratively to share information about how to detect the large and growing threat of terrorist content online. And they have improved their content-moderation systems to adapt to the bad actors' evolving techniques. These advancements have also improved other aspects of online services' moderation for separate kinds of harmful content.

Given this scale and complexity, effective content moderation is costly. Online services of all sizes engage in some content moderation, if only to comply with their legal obligations. Leading online services have spent billions of dollars refining their efforts at addressing harmful content. This requires investments into both human moderators capable of parsing wide permutations of communication as well as advanced technological means of detecting harmful content. Smaller online services, too, devote significant resources to content moderation—

potentially millions of dollars comprising large percentages of their annual revenues.

Even with significant investments into content moderation, online services' efforts will not be perfect, nor will there always be "correct" answers. When faced with billions of pieces of content—each with its own specific context—online services will make mistakes. The most well-intentioned content moderation will invariably result in occasional failures to preemptively remove harmful content and inadvertent removals of policy-compliant content. And when there are no "correct" answers, online services' users and advertisers may nevertheless disagree with the results of judgment calls.

The difficulties that individual pieces of content can raise for online services' content moderation is a crucial part of the process: online services have been able to learn from these instances to improve their content-moderation systems. These improved methods have produced the flourishing Internet we know today.

ARGUMENT

I. Online services foster communities for their users by implementing editorial policies about whether, and how, a service disseminates expression.

From the advent of the Internet, individuals and organizations set out to create unique communities by establishing their own online services. Those individuals and organizations have done so by implementing editorial policies delineating what expression will be acceptable for their online communities. Online communities can

dedicate themselves to specific sports teams,[5] popular culture interests,[6] intellectual pursuits,[7] religion,[8] and all manner of other shared interests and topics of debate.[9]

Regardless of their particular focus, online communities share at least one important goal: their users do not want to "com[e] across graphic photos and videos"—or harassment, scams, and other harmful material—"without warning."[10] Even for online services that wish to remove as little content as possible, "nudes, slurs, and criminal activity [have] forced administrators to rethink their entire content model."[11]

Thus, to foster their particular communities—and to comply with the law—online services *must* "monitor and moderate content on their sites because they recognize the potential for problematic content to appear that might contradict their values, undermine the trust of their other users, or threaten their ability to grow."[12] To ensure that

---

[5] *E.g.*, HuskerBoard.com Rules, HuskerBoard.com, https://bit.ly/3TWl8Gg ("Huskerboard welcomes any and every member that would like to discuss the Huskers.").

[6] *E.g.*, The Trek BBS, https://bit.ly/3OiEZOX (Star Trek forum).

[7] *E.g.*, Physics Discussion Forum, https://bit.ly/3OiCQTj.

[8] *E.g.*, Catholic Community Forum, https://bit.ly/3EpWg4h.

[9] There are also online services that have created broad communities, in which people can seek out different sub-communities. *E.g.*, Reddit Content Policy, Reddit, https://bit.ly/39bleIo ("Reddit is a vast network of communities that are created, run, and populated by you, the Reddit users. . . . Communities should create a sense of belonging for their members, not try to diminish it for others.").

[10] By the Numbers at 3.

[11] Joshua Luke Johnson, *A (Brief) History of Content Moderation*, Vocal, https://bit.ly/3Gz11uP.

[12] Engine, Startups, Content Moderation, & Section 230 2 (2021), https://bit.ly/3FS1HeB [hereinafter "Startup Report"].

their users understand what is permissible in their communities, online services require their users to "accept [their] designated guidelines regarding what types of accounts and posts are allowed on their sites."[13] This also allows users to select the online services that will offer them the communities they seek: online communities exist for every point of view, so content that some online services remove could be content that forms the core of other online services' communities.

What content online services deem harmful will vary according to the goals of their respective services and their users. For example, a forum dedicated to veganism may decide that it "will not tolerate members who promote contrary agendas" to the "vegan lifestyle" the website "promote[s]."[14] More generally, other services require simply that users "[s]tay on topic."[15]

But some categories of harmful content are widely prohibited across online services. At the most basic level, online services must remove certain content that violates the law,[16] such as child sexual abuse material.[17] These

---

[13] By the Numbers at 2.

[14] Terms and Rules, Vegan Forum, https://bit.ly/3TSeroT.

[15] LetsRun.com Message Board Community Guidelines, LetsRun.com, https://bit.ly/3USA75A [hereinafter "LetsRun Guidelines"].

[16] By the Numbers at 2.

[17] *See* 18 U.S.C. § 2258A; *see also, e.g.*, Community Guidelines, Parler, https://bit.ly/3OkbDQ2 ("Obvious examples include: child sexual abuse material, content posted by or on behalf of terrorist organizations, intellectual property theft.") [hereinafter "Parler Community Guidelines"].

efforts are invaluable to law enforcement nationwide, which often lacks the resources to track all of the potentially illegal content on the Internet.

In addition to illegal speech, online services remove "lawful but awful" speech that they find counterproductive to their communities. Many online services of all sizes prohibit pornography,[18] harassing or abusing other users,[19] hate speech,[20] and spam.[21] Particularly relevant to

---

And online services may prohibit further kinds of material that this Court has held protected by the First Amendment. *Ashcroft v. Free Speech Coal.*, 535 U.S. 234, 239 (2002) (addressing federal law prohibiting "sexually explicit images that appear to depict minors but were produced without using any real children"); *see* Our Child Sexual Exploitation Policy, Truth Social, https://bit.ly/3TP0yaT ("We have no tolerance towards any material or content that features or promotes child sexual exploitation in any way.").

Forms of illegal content present more difficulties. For example, videos "showing Isis recruitment can violate the law in one context, but also be legal and important for purposes such as documenting crimes for future prosecution." The Center for Internet & Society, YouTube Keeps Deleting Evidence of Syrian Chemical Weapon Attacks (June 26, 2018), https://bit.ly/3UUwLPD.

[18] *E.g.*, Terms of Use, Forumotion.com, https://bit.ly/3V3OQdK ("Content of a sexual, obscene, pornographic nature").

[19] *E.g.*, Forum Code of Conduct, Blizzard, https://bit.ly/3EIhmMv ("Causing disturbances in forum threads, such as picking fights, making off topic posts that ruin the thread, insulting other posters").

[20] *E.g.*, Frequently Asked Questions, Ubuntu, https://bit.ly/3TUeQXI ("Don't post anything that a reasonable person would consider offensive, abusive, or hate speech.").

[21] *E.g.*, Forum Guidelines, About Smoking Cessation Forum, https://bit.ly/3USsr31 ("We will delete spam/commercial posts.").

this case, online services prohibit various forms of pro-terrorist expression even if such speech is lawful.[22]

Beyond their commitments to protecting their communities, these services face a financial imperative to

---

[22] *E.g.*, Dangerous Individuals and Organizations, Meta, https://bit.ly/3nGnMBS ("content that praises, substantively supports, or represents events that Facebook designates as violating violent events - including terrorist attacks"); Community Guidelines, TikTok, https://bit.ly/3EfWYlJ ("organizations or individuals . . . who promote or engage in violence, including terrorist organizations"); Violent Organizations Policy, Twitter, https://bit.ly/3Tgs2WB ("affiliate with and promote the illicit activities of a terrorist organization or violent extremist group"—including by "providing or distributing . . . media/propaganda"); Violent Extremist or Criminal Organizations Policy, YouTube, https://bit.ly/3BWNMhx (expression "intended to praise, promote, or aid" terrorist organizations, including "[c]ontent praising or memorializing prominent terrorist . . . figures") [hereinafter "YouTube Violence Policy"]; Community Guidelines, Pinterest, https://bit.ly/3CANpKQ ("content and accounts that encourage, praise, promote, or provide aid to dangerous actors or groups and their activities" like "[t]errorist organizations"); Community Guidelines, Snap Inc., https://bit.ly/3WNC3xF ("no tolerance for content that advocates or advances violent extremism or terrorism"); Professional Community Policies, LinkedIn, https://bit.ly/3hpdIOi ("[c]ontent that depicts terrorist activity, that is intended to recruit for terrorist organizations, [and] that threatens, promotes, or supports terrorism in any manner"); WhatsApp Business Messaging Policy, WhatsApp, https://bit.ly/3FVwQxL ("organizations and/or individuals engaged in terrorist or organized criminal activity"); Do Not Post Violent Content, Reddit, https://bit.ly/3zVc4KU ("[t]errorist content, including propaganda"); Community Guidelines, Twitch, https://bit.ly/3thJ6B6 ("content that depicts, glorifies, encourages, or supports terrorism, or violent extremist actors or acts") (emphasis omitted); Community Guidelines, Truth Social, https://bit.ly/3USfTZs ("content posted by or on behalf of terrorist organizations"); Parler Community Guidelines ("content posted by or on behalf of terrorist organizations").

moderate content diligently. The market reacts swiftly when it perceives online services to have failed to sufficiently moderate objectionable and dangerous content. Specifically, users and advertisers have boycotted online services under such circumstances.[23] For example, in 2017, YouTube's advertisers boycotted YouTube and Google's other services after YouTube ran advertisements next to "objectionable content."[24] Similarly, in 2020, many of the world's largest companies boycotted Facebook "in a growing protest over how" Facebook handled "harmful content."[25] In the last few months, advertisers have also publicly paused placing their advertisements on

---

[23] To determine whether advertising on a particular online service would be detrimental to their brands, advertisers rely on their own judgments as well as third-party reports. For instance, the Global Alliance for Responsible Media publishes an annual report collecting data about online services' content moderation. *See generally* Global Alliance for Responsible Media (GARM), GARM Aggregated Measurement Report (Nov. 2022), https://bit.ly/3Fpqciy.

[24] Jack Nicas, *Google's YouTube Has Continued Showing Brands' Ads With Racist and Other Objectionable Videos*, Wall St. J. (Mar. 24, 2017), https://bit.ly/3fQC8jj.

In 2019, YouTube faced a similar boycott when "major brands" stopped purchasing advertisement space on YouTube after "their ads appeared on children's videos where pedophiles had infiltrated the comment sections." Daisuke Wakabayashi & Sapna Maheshwari, *Advertisers Boycott YouTube After Pedophiles Swarm Comments on Videos of Children*, N.Y. Times (Feb. 20, 2019), https://bit.ly/3WLP62m.

[25] Shannon Bond, *Over 400 Advertisers Hit Pause On Facebook, Threatening $70 Billion Juggernaut*, NPR (July 1, 2020), https://bit.ly/3TnNPfc.

Twitter over uncertainty about Twitter's content moderation.[26]

These boycotts confirm that users and advertisers will not support online communities riddled with harmful content. In part because of these clear signals from users and advertisers, content moderation has become a necessary component of operating any online service today.

II. Online services engage in careful, costly, iterative processes to improve content moderation—although mistakes and difficult judgment calls will be made given the vast amounts of expression online.

As the Internet has evolved, content moderation has become an increasingly large and complicated part of online services' operations. The vast scale of harmful content on the Internet requires constant vigilance from online services. Content moderation, therefore, requires significant expenditure of time and resources—largely because it is an iterative process of continually refining efforts at removing, and reducing the spread of, harmful and objectionable content. Given the immense scale of content on the Internet, it is impossible for content moderation to remain free of mistakes—or free of difficult judgment calls for which there is no "correct" answer.

---

[26] *Factbox: Advertisers React to Twitter's New Ownership*, Reuters (Nov. 18, 2022), https://bit.ly/3XgSQJw; *see also* Financial Services Company, *I Told My Team to Pause Our $750K/Month Twitter Ads Budget Last Week*, Blind, https://bit.ly/3AZ7iwv ("Our organic social and CS teams got dozens of screenshots of our ads next to awful content. Replies to our posts with hardcore antisemitism and adult spam remained up for days even when flagged.").

A. The scale of harmful content on the Internet and online services is immense.

The Internet contains an immense amount of expressive content reflecting all facets of humanity. Some estimate that people across the world create over 2.5 quintillion bytes of new data every day.[27] Placed in digestible context, this means that people upload millions of pieces of content across multiple online services every minute.[28] In that sea of expression, even a small percentage of harmful content reflects a large threat to all online services.

Accordingly, online services remove many individual pieces of harmful and objectionable content.[29] For example, a handful of online services removed nearly six billion posts in six months, from July to December 2020.[30] This includes approximately 45 million posts with graphic violence, 52 million posts with child sexual exploitation, 65.6

---

[27] Bernard Marr, *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*, Forbes (May 21, 2018), https://bit.ly/3glbqja.

[28] *Id.*

[29] Content removal is only one content-moderation tool that online services use. Services' policies identify a broad range of objectionable expression, but the harm from some expression is not so great that it requires removal and can be mitigated through other means like disclaimers, addenda, and prioritizing other content. *E.g.*, Twitter, Twitter 2.0: Our Continued Commitment to the Public Conversation (Nov. 30, 2022), https://bit.ly/3VHr0Va ("Our approach to policy enforcement will rely more heavily on de-amplification of violative content: freedom of speech, but not freedom of reach.") [hereinafter "Twitter 2.0"].

[30] By the Numbers at 1.

million posts with hateful content, and 25 million posts with abuse or harassment.[31]

An additional *three billion* removals were for spam.[32] Removing spam improves both user safety and user experience. Spam can be a source of malware, scams, and other fraud, so spam removal mitigates risks to users.[33] Furthermore, spam clutters users' feeds with unnecessary and irrelevant content, drowning out the other content.

Content moderation is therefore quite costly. It requires online services to expend time, money, and human capital. Even small startups with few users engage in at least some moderation to comply with their legal obligations. Indeed, startups that fail to provide their users with desirable communities will never grow. And as businesses grow in size and scope, so do the costs and scope of moderation.

Unsurprisingly, the country's largest online services expend enormous sums on content moderation. Some services have hired tens of thousands of employees and spent billions of dollars to develop and maintain their current content-moderation systems.[34]

Though smaller businesses face lower total costs, they often spend larger shares of their revenue and more money per user on content moderation. A recent survey of startups and mid-sized online services tracked how the

---

[31] *Id.*

[32] *Id.* at 1, 4.

[33] *Id.* at 4.

[34] *See* Meta, Our Progress Addressing Challenges and Innovating Responsibly (Sept. 21, 2021), https://bit.ly/3DTw8ym ("[W]e've started using technology that understands the same concept in multiple languages—and applies learnings from one language to improve its performance in others.") [hereinafter "Meta's Progress"].

costs of content moderation increase as businesses grow.[35] This survey identified startups as businesses with less than $100,000 in annual revenue that "serve between 1,000-5,000 monthly active users."[36] At that size, startup businesses often "do not yet encounter problematic content at a rate that requires a large moderation team or expensive, sophisticated moderation technology."[37] Consequently, most startups were able to rely on human content moderation, with training that could cost up to $10,000 annually.[38] Of those startups that used moderation technology, they spent between $40,000 and $1 million to develop or license the technology—with up to $50,000 per year in annual maintenance.[39] In total, content moderation can cost more than those startups generate in annual revenue.[40] And "the cost per user of each startups' moderation efforts ranged from a few dollars to over $150 per user."[41]

As startups become mid-sized firms, the resources they devote to content moderation increase dramatically—both in terms of spending and human capital.[42] In

---

[35] *See generally* Startup Report.
[36] *Id.* at 2. The surveyed startups ranged in size and age: "The companies are between 2 and 7 years old" and "have fewer than 10 employees." *Id.* Likewise, the mid-sized firms varied, and are "between 11 and 15 years old, generate more than $50 million to more than a billion dollars in annual revenue, employ 100 to 5000, and serve almost a million to just under a half-billion monthly active users." *Id.* at 3.
[37] *Id.* at 2.
[38] *Id.*
[39] *Id.*
[40] *Id.*
[41] *Id.*
[42] *Id.* at 3.

the same survey, all of the mid-sized companies employed human moderators (up to 250 total).[43] These companies spent between $1 million and $5 million "annually to retain their moderators," who tend to work those positions for one to three years.[44] And these companies spent up to $250,000 on training for their human moderators.[45]

Mid-sized firms also spent more on automated moderation.[46] The same survey reported that all surveyed firms use proprietary technologies, with some also licensing third-party technology.[47] These proprietary technologies cost between $500,000 and $30 million to develop, requiring between 500 and 300,000 engineering work hours.[48] After developing the technology, firms spend between $250,000 and $5 million to maintain those proprietary technologies.[49] Like startups, mid-sized firms also license moderation technology, which can cost between $250,000 and $5 million annually; implementation costs another $50,000 to $2 million and requires between 200 and 2,000 engineering hours.[50] As addressed below at pp.21-24, however, simply devoting more resources to content

---

[43] *Id.*
[44] *Id.*
[45] *Id.* The moderators do not work in a silo—their interaction with other departments like legal and public relations "could represent up to 10,000 work hours annually." *Id.*
[46] Because mid-sized firms generally have far more users than startups, their sharply increased *total* spending on moderation still results in a relatively low *per-user* cost: "as little as a number of cents to a few dollars." *Id.*
[47] *Id.*
[48] *Id.*
[49] *Id.*
[50] *Id.*

moderation will not solve content moderation's inherent difficulty.

B. Content moderation is an imperfect, iterative, and collaborative process, requiring constant refinement to address evolving threats.

While online services expend enormous resources improving their content-moderation efforts, harmful content constantly evolves and requires a dynamic response from online services. Operating any kind of online service "requires establishing rules and guidelines to thrive," but "[m]oderation is hard—really hard."[51] It is not enough for online services to develop content-moderation systems. Online services need to "iterate on [their] tactics because . . . bad actors will continue to change theirs."[52]

1. History demonstrates how, as the Internet has grown, content moderation has grown more complex along with it.

In the early days of the Internet, text-based web forums that served smaller communities were able to moderate content with minimal investments into human or technological resources. They could rely on simple technological screening and human moderators—often forum users acting as volunteers—to flag and remove harmful content. Those systems were sufficient for an age when the total volume of content on the Internet was smaller,

---

[51] LetsRun Guidelines.

[52] Meta, Combating Hate and Extremism (Sept. 17, 2019), https://bit.ly/3WDD1fE [hereinafter "Meta Combating Hate"]; *accord* Twitter 2.0 ("As we improve our policies and processes, bad actors will also develop new methods of disruption. This is not new. Our team of experts is constantly adapting to identify and defuse threats[.]").

websites were simpler, videos and sound were harder to easily disseminate over the Internet, and malicious actors were less sophisticated.

Since then, the volume of online content has exploded, and there have been countless innovations. Cheap and easy video livestreaming has exacerbated the challenges. Thus, online services moderate countless pieces of text, images, and video content every day.

Similarly, many modern online services have increasingly engaged in more sophisticated dissemination of information, using sorting mechanisms and recommendations to provide users with what the services believe will be most useful and informative. As the volume of content grows, users need features like search, sorting, and recommendation functions to see the content they find most valuable. These functions also allow online services to use more nuanced methods of content moderation, like deprioritizing borderline harmful content relative to other content.

In short, content moderation today requires more than a handful of volunteer moderators and simple technological screening techniques. Rather, many modern online services rely on "a combination of artificial intelligence and human reviewers [to] take down violating content as quickly as possible after it was posted."[53] Online

---

[53] By the Numbers at 5; *see also* Susan Wojcicki, Expanding Our Work Against Abuse of Our Platform, YouTube Official Blog (Dec. 5, 2017), https://bit.ly/3fT3cOX (YouTube employed more human reviewers, who "manually reviewed nearly 2 million videos for violent extremist content, helping train [YouTube's] machine-learning technology to identify similar videos in the future.") [hereinafter

services aim to engage in such proactive moderation before even a single user views the content.[54]

Algorithmic moderation enables proactive moderation. It effectuates the online services' human judgments about acceptable expression. And thus those algorithms rely on human moderation to train their algorithmic systems and to ensure that the algorithmic systems are working.[55]

2. Even with algorithmic help, addressing harmful content requires "constant effort and vigilance."[56] Content moderation at any scale is difficult because it depends on subjective judgments about human expression *in context*.

Whether an online service deems certain expression objectionable will often depend on the specific context of individual pieces of expression. Algorithmic moderation identifies precise text, images, or even sounds (as addressed below at pp.27-29). But it is much more difficult for algorithmic moderation to enforce nuanced policies

---

"YouTube's Expanding Work"]; *accord* Examining Social Media Companies' Efforts to Counter Online Terror Content and Misinformation: Hearing Before the H. Comm. on Homeland Sec., 116th Cong. (2019) (statement of Derek Slater, Global Director of Information Policy, Google), https://bit.ly/3G1hI1V (In the first quarter of 2019, "YouTube manually reviewed over" 1 million "suspected terrorist videos and found that only fewer than 10% (90 [thousand] videos) violated [YouTube's] terrorism policy.").

[54] As discussed below (at pp.21-24), regardless of when an online service is alerted to harmful content, content moderation is always complex.

[55] *See* YouTube's Expanding Work.

[56] By the Numbers at 1.

that require contextual considerations.[57] That requires online services to make judgment calls.

There are myriad examples of these difficulties, ranging from identifying and moderating new, trendy forms of dangerous content to successfully moderating the ever-present problems of pro-terrorist or hateful content.

The "Tide pod challenge" illustrates this problem well. A few years ago, it became trendy for people to upload videos of themselves and others eating (or pretending to eat) detergent pods to various online services.[58] Because consuming laundry detergent is dangerous, people engaging in this "challenge"—whether they uploaded videos or not—deluged poison control hotlines.[59] Of particular concern, this dangerous trend was popular among children and teenagers.[60]

The online services took action.[61] For instance, YouTube concluded that Tide-pod-challenge videos violated YouTube's already-existing policy against "content that's intended to encourage dangerous activities that have an inherent risk of physical harm."[62] And thus

---

[57] *E.g.*, Twitter 2.0 ("[W]e will make mistakes, we will learn, and we will also get things right.").

[58] Brian Feldman, *YouTube Is Pulling Videos Featuring the Tide Pod Challenge*, N.Y. Mag. (Jan. 18, 2018), https://bit.ly/3umbUsO [hereinafter "Feldman, *Tide Pod Challenge*"].

[59] Christopher Ingraham, *There Were Over 12,000 Poison Control Calls for People Eating Laundry Pods Last Year*, Wash. Post (Jan. 16, 2018), https://bit.ly/3Vypg13.

[60] Michelle Toh, *Tide Pod Challenge: YouTube Is Removing 'Dangerous' Videos*, CNN Business (Jan. 18, 2018), https://bit.ly/3F04jEY.

[61] *Id.*

[62] *Id.*; *see* Harmful or Dangerous Content Policies, YouTube, https://bit.ly/3Fm6yUI.

YouTube could remove such videos consistent with its policies. But YouTube did not flatly prohibit any and all mentions of the Tide pod challenge. Instead, it "allowed" "videos commenting or reporting on the trend."[63] It is not always easy, however, to distinguish between videos engaging in (or advocating for) the Tide pod challenge, and videos parodying or criticizing the Tide pod challenge. Malicious actors who are aware that online services will permit "parody" and "criticism" will adjust their presentation of harmful content accordingly. Algorithmic moderation, despite continual advancements, will not always parse this correctly. Nor will there always be "correct" answers that *any* moderation system could identify; there will always be difficult judgment calls.

The difficulties posed by the Tide pod challenge—just *one* form of trendy content—are illustrative of the countless content-moderation challenges that online services face. For instance, online services have had difficulty removing content that glorifies Nazis and Nazi ideology.[64] In efforts to remove such content, those online services might mistakenly take "down content that educate[s] and inform[s] people about Nazis and their ideology."[65] This

---

[63] Feldman, *Tide Pod Challenge*.

[64] YouTube has a policy that expressly permits certain harmful content if it is "related to terrorism or crime for an educational, documentary, scientific, or artistic purpose" and "provide[s] enough information in the video or audio itself so viewers understand the context." YouTube Violence Policy. For instance, a documentary explaining the horrors of the Holocaust that includes part of a speech by Adolf Hitler could be treated differently than the same speech uploaded by itself.

[65] Trust & Safety Foundation, YouTube's New Policy on Nazi Content Results in Removal of Historical and Education Videos (2019), https://bit.ly/3UDqirt.

includes content with "archival footage of propaganda speeches by Nazi leaders, including [Adolf] Hitler."[66] Whether a piece of content glorifies Nazis—as compared to educating about the Nazis' atrocities—is a case-by-case, context-specific inquiry.

As another illustration, to remove depictions of the Syrian conflict's atrocities, YouTube removed "documentation of chemical attacks, attacks on hospitals and medical facilities, [and] destruction of civilian infrastructure."[67] But the Syrian Archive and other groups have relied on user-generated videos to investigate human rights violations. As a result, after these videos were removed under YouTube's moderation policies, the Syrian Archive "raised this issue with YouTube," and "over 650,000 videos were restored and made publicly viewable since July 2017."[68]

3. Online services' efforts to remove terrorist content demonstrate how far they have gone to minimize the amount of harmful content online—as well as their successes in addressing this form of dangerous content.

Online services face similar threats of pro-terrorist speech. This has prompted industry-wide collaboration. Specifically, many online services are members of, or benefit from, organizations like Tech Against Terrorism and the Global Internet Forum to Counter Terrorism. These

---

[66] *Id.*; *see also id.* ("Another YouTube user noticed his upload of 1938 documentary about the rise of the Nazi party in Germany had been taken down for similar reasons, even though the documentary was decidedly anti-Nazi in its presentation and had obvious historical value.").

[67] Syrian Archive, Removals of Syrian Human Rights Content (May 2020), https://bit.ly/3DTGPAT.

[68] *Id.*

are organizations that connect online services, government, and other counter-terrorism experts, while maintaining archives of terrorist content.[69]

This collaboration has proven successful. From December 2020 to November 2021, Tech Against Terrorism's "intelligence experts submitted 18,958 URLs containing terrorist content, and the TCAP [Terrorist Content Analytics Platform] sent 11,074 alerts to 65 tech companies" and 94% of that content "is now offline."[70] Now, more than 114 companies receive such alerts from Tech Against Terrorism.[71] Online services like "photo sharing, video hosting, and audio streaming services, as well as web hosting platforms, are most responsive" to these alerts "and have removed 100% of verified terrorist content notified via the TCAP."[72]

To reap the benefits of that collaboration, online services have also developed their own processes to remove harmful content. Individually, online services have had success removing terrorist content, ideally before anyone accesses that content. Among the companies that report their success in removing terrorist content,[73] online

---

[69] About Tech Against Terrorism, Tech Against Terrorism, https://bit.ly/3Vs7YTa; Global Internet Forum to Counter Terrorism, Annual Report 10-17 (December 2021), https://bit.ly/3Uq7kUM.

[70] Tech Against Terrorism, Terrorist Content Analytics Platform Year One: 1 December 2020 – 30 November 2021 2 (Mar. 17, 2022), https://bit.ly/3Tw3FEI.

[71] *Id.*

[72] *Id.* at 3.

[73] The OECD has noted that online services' transparency and reporting of their anti-terrorism efforts has improved significantly in recent years. *See* OECD, Transparency Reporting on Terrorist and Violent

services report increasing success in proactively removing terrorist content.[74] Leading U.S. online services report proactively removing between 90% and 99% of such content.[75]

That success is partially the result of advanced algorithmic moderation designed to address multiple kinds of media in many languages. Online services use what is sometimes called "hashing" (or "media-matching" or "content-matching") technology: "convert[ing] a file into a unique string of digits that serves as a 'fingerprint' of that file."[76] In other words, online services can identify specific pieces of media and restrict their upload or

---

Extremist Content Online 5 (July 2021), https://bit.ly/3Y2NdPt ("[T]he degree of transparency and clarity in the top 50 services' TVEC-related [terrorist and violent extremist content] policies and procedures has improved appreciably.").

[74] *E.g.*, Jennifer O'Connor, Building Greater Transparency and Accountability with the Violative View Rate, YouTube Official Blog (Apr. 6, 2021), https://bit.ly/38noixm (YouTube's "violative view rate" has decreased 70% since 2017).

[75] *E.g.*, Meta, Facebook Community Standards Enforcement Report – Dangerous Organizations: Terrorism and Organized Hate, https://bit.ly/3Thc9iF; Meta, Instagram Community Standards Enforcement Report – Dangerous Organizations: Terrorism and Organized Hate, https://bit.ly/3DQ8P8w; Twitter, Rules Enforcement (July 28, 2022), https://bit.ly/3tdZRgt; Google, YouTube Transparency Report – Featured Policies, https://bit.ly/3FYqfT0 [hereinafter "YouTube 2022 Report"]; Twitch, H2 2021 Transparency Report, https://bit.ly/3NQO8xK.

[76] Dangerous Organizations: Terrorism and Organized Hate, Meta, https://bit.ly/3DQ8P8w; Examining Social Media Companies' Efforts to Counter Online Terror Content and Misinformation: Hearing Before the H. Comm. on Homeland Sec., 116th Cong. (2019) (statement of Monika Bickert, Head of Global Policy Management, Facebook), https://bit.ly/3hovam3; YouTube 2022 Report.

otherwise remove them immediately. Services that have "long used image- and video-hashing" have also added "audio- and text-hashing techniques for detecting terrorist content."[77] Likewise, online services have trained their algorithmic moderation to "identify violating text posts" in 19 languages.[78]

But hashing technology is not a cure-all. Online services can only hash and "identify copies of *known* bad material"—not necessarily new forms of harmful content.[79] For every video that online services can effectively hash, malicious actors can create infinite copies of that video with enough differences to avoid immediate algorithmic moderation. Thus, online services are constantly on the lookout for new pieces of harmful content.

4. One recent example illustrates the difficulties that a single piece of new harmful content can pose to online services—even when a service successfully removes the original content quickly.

In 2019, a terrorist murdered 51 people and injured 40 more at a mass shooting at two mosques in Christchurch, New Zealand. The terrorist livestreamed parts of his attack by video.[80] The video was removed "within minutes" of the online service being informed by local police about the video.[81] Nevertheless, that *single* live-streamed video on *one service* has spawned countless copies of the video

---

[77] Bickert Testimony at 3.
[78] *Id.*; *accord* Meta's Progress.
[79] Meta Combating Hate (emphasis added).
[80] Meta, Update on New Zealand (Mar. 18, 2019), https://bit.ly/3hrgyCv [hereinafter "Meta's Update on New Zealand"].
[81] *Id.*

across the Internet. It remains a constant target of content-moderation enforcement to this day.[82]

Removing the original livestreamed video was just the beginning. At the time the original video was removed, the live broadcast had been viewed 200 times and the original video had been viewed 4,000 times.[83] Nevertheless, copies of that video proliferated. Despite the limited reach of the original video, one service alone "removed about 1.5 million videos of the attack globally" within the first 24 hours of the terrorist attack.[84] That service was able to "block[]" 1.2 million of those videos "at upload"—which prevented users from seeing those versions of the video.[85] Other services had to adapt and address this new video with all its variants.[86]

The Christchurch video illuminated areas of improvement for content moderation at that time. Notably, the Christchurch video did not prompt "automatic detection systems" because those systems "did not have enough content depicting first-person footage of violent events to effectively train [] machine learning technology."[87] Since then, however, companies have worked with law enforcement in multiple countries to obtain first-person footage, which can "train" their systems to distinguish videos like

---

[82] *E.g.*, Nikki Main, *Twitter Failed to Flag Christchurch Shooting Video Re-Uploads*, Gizmodo (Nov. 28, 2022), https:/bit.ly/3Hcm9HM.
[83] Meta's Update on New Zealand.
[84] *Id.*
[85] *Id.*
[86] *See* Jane Wakefield, *Christchurch Shootings: Social Media Races to Stop Attack Footage*, BBC (Mar. 16, 2019), https://bbc.in/3EfPhfl.
[87] Meta Combating Hate.

the one of the Christchurch attack from "fictional content from movies or video games."[88]

Experience with the Christchurch video has led to additional improvements in content moderation. For example, the continued spread of copies of the Christchurch video has prompted further improvements to hashing technology—such that "other shares that are visually similar to that video are then detected and automatically removed."[89] Just four days after the attack, one online service identified "more than 800 visually-distinct videos," and there was evidence that malicious actors were using other websites to circumvent content moderation efforts.[90] In particular, "screen recordings" of the video proved "more difficult to detect," so (as described above at pp.26-27) the service developed "additional detection systems including the use of audio technology" to automatically detect videos.[91]

Online services have learned from the Christchurch video. Unfortunately, malfeasors still attempt to livestream their atrocities. But online services have improved their ability to stop and remove those videos more quickly. For example, earlier this year, another attacker

---

[88] *Id.* The difference between fictional or simulated first-person violence and livestreamed crimes underscores the importance for content moderation to evaluate the context of content.

[89] Meta's Update on New Zealand.

[90] *Id.*

Six months later, the online service had identified 900 "visually unique versions of this video that it [could] automatically detect and remove." Olivia Solon, *Six Months After Christchurch Shootings, Videos of Attack Are Still on Facebook*, NBC News (Sept. 20, 2019), https://bit.ly/3VWmMsP.

[91] Meta's Update on New Zealand.

in Buffalo livestreamed a racist mass-shooting on another online service.[92] That service was able to "shut down" the video in "less than 2 minutes after the violence began."[93] But this video also spread around the Internet and remains a constant source of moderation difficulties.

* * *

As malicious actors evolve, so too do online services' content-moderation efforts. The same technological developments that help shape communities on the Internet can also be used to undermine those communities. Online services preserve their communities by responding aggressively to evolving threats. Content-moderation efforts have not always been, and cannot be, perfect. But even these imperfect efforts are an important facet of operating modern online services, which all face constant threats from malicious actors and harmful content. Without such moderation, dangerous and objectionable content accessible on online services would proliferate exponentially—and the communities that define the Internet would wither.

CONCLUSION

The judgment of the court of appeals should be reversed.

---

[92] Emma Bowman, Bill Chappell, & Becky Sullivan, *What We Know So Far About the Buffalo Mass Shooting*, NPR (May 16, 2022), https://bit.ly/3HadJQX.

[93] *Id.*

Respectfully submitted.

Scott A. Keller
Counsel of Record
Steven P. Lehotsky
Jeremy Evan Maltz
LEHOTSKY KELLER LLP
200 Massachusetts Ave., NW
Washington, DC 20001
(512) 693-8350
scott@lehotskykeller.com

Counsel for Amici Curiae

DECEMBER 2022