UCLA

JOURNAL OF LAW & TECHNOLOGY

*SPECIAL ISSUE: GOVERNING THE DIGITAL SPACE*

# THE RIGHTS AND WRONGS OF FOLK BELIEFS ABOUT SPEECH: IMPLICATIONS FOR CONTENT MODERATION

Aileen Nielsen*

ABSTRACT

Political discourse and survey research both suggest that many Americans believe constitutional protections for free expression extend more broadly than what is reflected in the black letter law. A notable example of this has been the claim—sometimes explicitly constitutionalized—that content moderation undertaken by digital platforms infringes on users' legally protected freedom of expression. Such claims have proven both rhetorically powerful and politically durable. This suggests that laypeople's beliefs about the law—distinct from what the state of the law actually is—could prove important in whether content moderation policies are democratically and economically successful.

* Fellow in Law & Tech at the Center for Law and Economics at ETH Zurich. She received her J.D. from Yale Law School and her A.B. from Princeton University. For helpful comments, she thanks Barbara Prainsack, Nikolaus Forgo, Laura Fichtner, Martin Riedl, Rachel Griffin, Naoki Matsuyama, and Paul Dunshirn, as well as participants at the University of Vienna 2022 Winter School, Taming the iMonster.

This Article presents the results of an experiment conducted on a large, representative sample of Americans to address questions raised by the phenomenon of constitutionalized rhetoric about digital platforms and content moderation. The experimental results show that commonly-held but inaccurately broad beliefs about the scope of First Amendment restrictions are linked to lower support for content moderation. Yet constitutional information presented to participants to correct such misapprehensions backfires, leading to lower support for content moderation. These results highlight an undertheorized difficulty of developing widely acceptable content moderation regimes, while also demonstrating a surprising outcome when correcting misrepresentations about the law.

TABLE OF CONTENTS

## INTRODUCTION

On January 8, 2021—two days after the 2021 United States Capitol attack—Twitter took the unusual step of permanently suspending then-President Donald Trump's Twitter account "due to the risk of further incitement of violence."[1] In doing so, the company referred both to its "public interest framework"[2] and to its "Glorification of Violence" policy,[3] providing the rationale for this unprecedented and highly controversial enforcement action by a digital platform of its own privately designed online speech policies. The company's actions, removing one of the most powerful people in the world and citing its own contractual policy documents to do so, reflected the legal realities expressed in black letter law—that digital platforms are free to undertake content moderation activities on the strength of their direct relationships with users as governed by typical sources of private law, particularly that of contract.

Yet, the decision was a watershed moment for discourse about public law and policy, generating an unprecedented level of discussion on social media about content moderation.[4] Many criticized Twitter's decision to suspend Trump's account. Some critiques included references to free

---

1. *Permanent Suspension of @realDonaldTrump*, TWITTER: BLOG (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.

2. *World Leaders on Twitter: Principles & Approach*, TWITTER: BLOG (Oct. 15, 2019), https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.

3. *Glorification of Violence Policy*, TWITTER: HELP CTR. (Mar. 2019), https://help.twitter.com/en/rules-and-policies/glorification-of-violence.

4. Meysam Alizadeh et al., *Content Moderation as a Political Issue: The Twitter Discourse Around Trump's Ban* 9 (Dep't of Pol. Sci., Univ. of Zurich, Working Paper, 2021), https://fabriziogilardi.org/resources/papers/content-moderation-twitter.pdf. The authors undertook a comprehensive analysis of all discussion of content moderation on Twitter, as identified by Twitter's academic API. *Id.* at 6–7. They found that "[t]he salience of content moderation was very low until Twitter started fact-checking Donald Trump's tweets in June 2020. After that, the issue only received a moderate amount of attention, as can be seen in the number of retweets. The peak was reached, unsurprisingly, when Donald Trump was initially banned from Twitter for incitement of violence in the context of the assault on the Capitol on January 6, 2021." *Id.* at 9.

speech,[5] although what free speech meant in the context was not always made clear.

There was, certainly, a constitutional component to the free speech discussion. Mainstream news media organizations, likely responding to ordinary Americans' upset and confusion in the days following the suspension, addressed explicitly the question of whether Trump's constitutional rights had been violated by Twitter.[6] Legal commentators cited in mainstream media coverage took the uncontroversial position that the suspension was a straightforward exercise of Twitter's own constitutionally protected right of free expression.[7] But political commentators of all persuasions criticized the degree of power exercised by Twitter—a private company—without democratic oversight.[8] Likewise, some in the judiciary and legal academia have questioned whether the powerful position or socially important functions of digital platforms might

---

5. *See, e.g.*, Adam Satariano, *After Barring Trump, Facebook and Twitter Face Scrutiny About Inaction Abroad*, N.Y. TIMES (Jan. 17, 2021), https://www.nytimes.com/2021/01/14/technology/trump-facebook-twitter.html (including the phrase "free speech" twice, once in a quote and once in the writer's own choice of words). Even news coverage that explicitly purports to address the "First Amendment" tends to use "free speech" interchangeably both in the reporter's own language and in the quotation by experts. *E.g.*, Adam Liptak, *Can Twitter Legally Bar Trump? The First Amendment Says Yes*, N.Y. TIMES (Jan. 9, 2021), https://www.nytimes.com/2021/01/09/us/first-amendment-free-speech.html ("'To take an account down in these circumstances is not an affront to free speech, as some have suggested,' Mr. Jaffer said. 'To the contrary, it's the responsible exercise of a First Amendment right.'").

6. Consider that many mainstream media articles discussed whether Trump's constitutional rights had been violated. *See, e.g.*, Lauren Giella, *Fact Check: Did Twitter Violate President Trump's First Amendment Rights?*, NEWSWEEK (Jan. 11, 2021, 7:38 PM), https://www.newsweek.com/fact-check-did-twitter-violate-president-trumps-first-amendment-rights-1560673.

7. *See, e.g.*, *id.*; *see also* Fred Hiatt, Opinion, *Legally, Trump's Tech Lawsuit Is a Joke. But It Raises a Serious Question*, WASH. POST (July 8, 2021, 6:23 PM), https://www.washingtonpost.com/opinions/legally-trumps-tech-lawsuit-is-a-joke-but-it-raises-a-serious-question/2021/07/08/33bc2dfa-e010-11eb-9f54-7eee10b5fcd2_story.html.

8. *See, e.g.*, Ryan Browne, *Germany's Merkel Hits Out at Twitter Over 'Problematic' Trump Ban*, CNBC (Jan. 11, 2021, 1:53 PM), https://www.cnbc.com/2021/01/11/germanys-merkel-hits-out-at-twitter-over-problematic-trump-ban.html.

render First Amendment restrictions applicable to such actors, or, alternatively, might limit the scope of their own First Amendment rights.[9]

Six months after the account suspension, Trump sued Twitter,[10] alleging infringements of his First Amendment rights. The lawsuit was judged by some observers as so clearly lacking in merit that the attorneys who filed the claim were thought to be at risk of sanctions for filing

---

9. For a "plausible (though far from open-and-shut) argument" that prohibiting platforms from engaging in some forms of content moderation would be constitutionally permissible and not violative of platforms' own rights of expression, see Eugene Volokh, *Treating Social Media Platforms Like Common Carriers?*, 1 J. FREE SPEECH L. 377, 414–52 (2021). It is also worth noting a concurrence issued in 2021 by Justice Thomas, wherein Justice Thomas speculated about potential legal theories that might limit the First Amendment freedoms of digital platforms. He also highlighted the urgency of the issue. "We will soon have no choice but to address how our legal doctrines apply to highly concentrated, privately owned information infrastructure such as digital platforms." Biden v. Knight First Amend. Inst. at Columbia Univ., 141 S. Ct. 1220, 1221 (2021). But neither appellate courts nor the Supreme Court have, thus far, held that a digital platform's activities or role or relation to government justified bounding the platform by the First Amendment. *See, e.g.*, Manhattan Comty. Access Corp. v. Halleck, 139 S. Ct. 1921 (2019); Prager Univ. v. Google LLC, 951 F.3d 991 (9th Cir. 2020).

These notions of potential regulation or First Amendment restrictions on the content moderation activities of digital platforms are not the only developing view. For example, in recent work, Olivier Sylvain describes "an emerging view that companies, especially internet companies, have a constitutional right to decide which ideas to distribute or promote and which ideas to demote or block." Olivier Sylvain, *Platform Realism, Informational Inequality, and Section 230 Reform*, 131 YALE L.J. 475, 496 (2021), https://www.yalelawjournal.org/forum/platform-realism-informational-inequality-and-section-230-reform. Further, Sylvain observes that—due to judicial interpretations of Section 230 of the Online Communications Decency Act—digital platforms may in fact receive more judicial deference than do traditional publishers. *Id.* at 494.

10. Complaint for Injunctive and Declaratory Relief, Trump v. Twitter, Inc., No. 1:21-cv-22441 (S.D. Fla. July 7, 2021), https://www.wsj.com/media/TrumpvTwitter.pdf. Trump argued that Twitter was bound by the First Amendment because "Defendant Twitter's status thus rises beyond that of a private company to that of a state actor, and as such, Defendant is constrained by the First Amendment right to free speech in the censorship decisions it makes." *Id.* at 2, ¶ 3. Trump's theory was premised, inter alia, on alleged coercion by Democratic legislators to force Twitter to censor Trump. *Id.* at 10–11, ¶¶ 48–61. It was also based on the theory that government legislation encouraged Twitter to censor content. *Id.* at 14–21, ¶¶ 62–91.

frivolous litigation.[11] Observers noted that the lawsuit failed basic procedural requirements (for example, it was not filed in the appropriate venue), and further that the suit was more likely to serve as a vehicle for fundraising than as an opportunity to adjudicate valid (or sincere) constitutional arguments.[12] Such comments were, presumably, based on the fact that it is generally uncontroversial that First Amendment restrictions do not apply to private entities.[13] Yet, Trump was not alone in his tenuous constitutional arguments. Some politicians have seemingly made a point of stoking the content moderation debate and constitutionalizing the terms on

---

11. *See, e.g.*, *Trumps Lawsuits, Defending Journalists & Gorsuch's Actions*, #SISTERSINLAW (July 10, 2021), https://www.jillwinebanks.com/blog/2021/7/10/trumps-lawsuits-defending-journalists-gorsuchs-actions.

12. *See, e.g.*, Roger Sollenberger, *Trump's Tech Lawsuit Already Turning into Fundraising Scheme*, DAILY BEAST (July 10, 2021, 3:39 AM), https://www.thedailybeast.com/donald-trumps-tech-lawsuit-already-turning-into-fundraising-scheme. For a similar observation made before the 2021 United States Capitol Attack, *see Deplatformed: Social Media Censorship and the First Amendment*, MAKE NO L. (Aug. 28, 2019), https://legaltalknetwork.com/podcasts/make-no-law/. More generally, misrepresentations about the state of the law as applied to content moderation seem to be deployed, possibly strategically, by other elected figures beyond Trump. Misrepresentation of the scope of First Amendment protections and other misrepresentations about the state of the law regarding online platforms and content moderation have been observed in use by other politicians. For example, in a 2018 Senate hearing, Senator Ted Cruz repeatedly questioned Facebook CEO Mark Zuckerberg regarding whether Facebook is a neutral public forum, implying (wrongly) that such neutrality is—or ought to be—necessary for protection under Section 230 of the Online Communications Decency Act. Leigh Beadon, *Ted Cruz Gets Section 230 All Wrong, While Zuck Claims He's Not Familiar with It*, TECHDIRT (Apr. 10, 2018, 3:29 PM), https://www.techdirt.com/articles/20180410/13530139604/ted-cruz-gets-section-230-all-wrong-while-zuck-claims-hes-not-familiar-with-it.shtml.

13. *See* Evelyn Douek, *Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV. 759, 768 (2021) ("[I]t is relatively uncontroversial that private actors can restrict more speech than governments."). Notable cases in which private entities have been limited from restricting the expression of others on their property include the company town case of *Marsh v. Alabama*, 326 U.S. 501 (1946), in which a company had taken on traditional governmental roles in running a company town, and the shopping mall case of *Pruneyard Shopping Ctr. v. Robins*, 447 U.S. 74 (1980), in which the United States Supreme Court found it permissible for the state of California to interpret its state constitution to protect political protesters from being evicted from private property when that property had been held open to the public.

which the conversation takes place.[14] While we cannot know the intent of any particular individual, it seems reasonable to infer that some politicians may strategically misrepresent the state of First Amendment law so as to create the false[15] belief or otherwise politically benefit from false belief that the First Amendment typically restricts private actors.[16]

It likewise seems plausible that digital platforms could suffer business losses due to reputational costs that could be particularly acute if many people believe that content moderation constitutes a constitutional infringement. In the days following Twitter's suspension of Trump's account, conservative media outlets and politicians were quick to allege that Twitter was losing followers.[17] Given that such information is

---

14. *See generally* Jeff Parrott, *Conservatives & Social Media: Are Free Speech Rights Being Violated?*, DESERET NEWS (July 1, 2021, 4:06 PM), https://www.deseret.com/indepth/2021/1/12/22225290/parler-amazon-facebook-twitter-conservatives-social-media-free-speech (providing examples of Republican politicians complaining about infringement of their speech rights by private actors). Parrott describes how Senator Josh Hawley called the decision by a private publisher not to publish his book a "direct assault on the First Amendment."). *Id.*; *see also, e.g.*, John Cornyn (@JohnCornyn), TWITTER (Jan. 11, 2021, 7:35 AM), https://twitter.com/JohnCornyn/status/1348654736791244807 (a Tweet by a current U.S. Senator reading "Are social media platforms 'de facto public squares?' I am inclined to think so.").

15. It is beyond the scope of this article to establish that beliefs that content platforms are bound by the First Amendment are necessarily inaccurate. For readers who find the use of "false beliefs" or "inaccurate" (this latter I will use throughout the remainder of this Article) to be problematic, these labels can be understood to mean "lacking precedent" or "tenuous", so as to reflect the lack of legal precedent for the idea of applying First Amendment restrictions to digital platforms. In any case, the term is not meant to be pejorative but simply descriptive.

16. For this experiment it is unimportant whether strategic politicians create false beliefs about the law or seek to exploit existing misapprehensions by laypeople (or both). This question would be an interesting and important topic for investigation on its own; however, it would be difficult to find or create an experimental scenario in which causality could be fairly inferred, given the difficulties of controlling exposure to such rhetoric.

17. For example, the New York Post claimed that "Trump fans ditch Twitter en masse after [Trump's] suspension," a claim that Trump himself also made. Jon Levine, *Trump Fans Ditch Twitter En Masse After President's Suspension*, N.Y. POST (Jan. 9, 2021, 10:09 AM), https://nypost.com/2021/01/09/trump-fans-leave-twitter-after-presidents-suspension/. It is worth noting, however, that Twitter's quarterly earnings figures did not

proprietary, the claim is difficult to verify. There was, however, a significant drop in Twitter's market capitalization in the days following the suspension.[18] From an ex ante perspective, the company could have reasonably feared significant business losses from the decision, especially if laypeople perceived Twitter as trampling upon the Constitution.[19] Indeed, the perception that Twitter violated the Constitution, while legally baseless, may be widely held. As will be further discussed in Part II, survey evidence shows that many Americans incorrectly believe that platforms' content moderation activities are an infringement of constitutional rights.[20]

More broadly, American society is conflicted with regard to what constitutes acceptable online speech governance. On the one hand, a large portion of Americans judge that online platforms have a responsibility to address offensive or otherwise problematic online content; on the other hand, a large portion of Americans distrust digital platforms to moderate

---

tend to support the assertion that the account suspension had done long-term harm. *See* Natalie Colarossi, *Trump Claims 'Boring' Twitter Is Losing Users as the Platform Gains Millions of Them*, NEWSWEEK (May 1, 2021, 10:23 AM), https://www.newsweek.com/trump-claims-boring-twitter-losing-users-platform-gains-millions-them-1588051 ("While Trump's statement that Twitter's stock prices have fallen is true, the company still recorded a healthy increase in new followers in its first-quarter earnings reported. More than 7 million new daily users joined the platform—up by 20 percent from a year ago—while ad revenue increased by 32 percent, according to CNBC.").

18. Ambar Warrick & Sruthi Shankar, *Twitter Tumbles as Trump Ban Puts Social Media in Spotlight*, REUTERS (Jan. 11, 2021, 4:01 AM), https://www.reuters.com/article/us-twitter-stocks-trump/twitter-tumbles-as-trump-ban-puts-social-media-in-spotlight-idUSKBN29G0XG.

19. It is of course also possible that substantial backlash could result for other reasons, including normative free speech concerns independent of the status of constitutional law or concerns consistent with the spirit of the First Amendment even if not required by black letter law. Such reactions are likely important in understanding concerns that were raised about Twitter's decision, but these motivations are not the subject of this experiment. Such concerns already seem reflected in the scholarly discourse and empirical survey work undertaken. This work looks to a distinct and underexplored factor: misapprehensions as to the state of the law.

20. Other recent work has more broadly identified a trend of low levels of constitutional knowledge. *See generally* Kevin L. Cope & Charles Crabtree, *Knowing the Law*, U. CHI. L. REV. ONLINE (Apr. 2020), https://lawreviewblog.uchicago.edu/2021/04/05/cv-cope-crabtree/.

content and even believe that actions taken by platforms to address problematic online speech violate the First Amendment. So digital platforms confront a divided society and thus, a divided consumer base. Platforms are simultaneously judged to be morally obligated to moderate content while also morally—or even constitutionally—obligated not to moderate content. The latter perceived obligation may result from an inaccurate understanding of First Amendment restrictions. Perhaps Americans who oppose content moderation practices do so, at least in part, because they (wrongly) believe such practices are unconstitutional. Perhaps, increased First Amendment literacy would reduce the degree to which platforms face conflicting expectations from the American public.

This Article describes an experimental investigation into this very question. Unlike most scholarship on the topic of content moderation,[21] this Article looks to what laypeople think about the state of the law—rather than plausible arguments made by legal academics[22] or controlling legal holdings crafted by judges[23]—to understand an important but

21. As has been highlighted by others, related scholarship tends to focus on conceptual issues rather than the opinions of laypeople. *See, e.g.*, Martin J. Riedl et al., *Antecedents of Support for Social Media Content Moderation and Platform Regulation: The Role of Presumed Effects on Self and Others*, INFO., COMMC'N & SOC'Y 2 (2021), https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.1874040 (opining that existing scholarship "often ignore[s] critical questions about what users and nonusers think.").

22. Volokh, *supra* note 9, at 414–52.

23. *See* Manhattan Comty. Access Corp. v. Halleck, 139 S. Ct. 1921 (2019). Although this case was not about a digital platform, it was widely understood to reaffirm legal precedents that narrowly limit the circumstances in which private actors are bound by the First Amendment. *See, e.g.*, *Manhattan Community Access Corp. v. Halleck*, 133 HARV. L. REV. 282 (2019), https://harvardlawreview.org/2019/11/manhattan-community-access-corp-v-halleck/ ("Put simply, opening up private property to others' speech does not turn the property into a public forum because an entity can open a public forum only if it is already a state actor. A contrary rule would strip private property owners of editorial liberties by subjecting them to the First Amendment whenever they opened their property for speech. The Court in Hudgens v. NLRB denied such a suggestion, and the Halleck Court reaffirmed that holding." (internal citations omitted)).

underexplored connection between Constitutional law and private firms' policies regarding speech on their platforms.[24]

This Article starts from the normative perspective that some level of content moderation is desirable and even necessary for the future of online content. It is therefore both interesting and important to understand whether digital platforms that seek to promote content moderation, or legislators who seek to enact legal reform on this issue, could face particular challenges arising from inaccurate perceptions about constitutional law.

This Article makes no further normative assumptions about the ideal content moderation regime or whether the current status of online speech regulation is desirable or lamentable. Rather, this Article looks to expand notions about what is relevant when exploring novel solutions for online speech regulation, specifically the beliefs and opinions of the ordinary people who constitute the user base of affected firms and the constituents of lawmakers likely to be drafting future relevant policy initiatives.

This Article proceeds in four Parts. In Part I, this Article offers a brief overview of recent survey work that strongly suggests both that laypeople misapprehend the state of First Amendment law and that there is a connection between such misapprehensions and attitudes about content moderation. In Part II, this Article discusses the relevance of an educational intervention to the topic at hand. In Part III, this Article describes a vignette experiment to measure the relationship between constitutional misapprehensions and opinions about content moderation. In Part IV, the Article presents experimental results demonstrating a relationship between constitutional misapprehensions and opinions about content moderation, but also, a backfire effect whereby constitutional information to correct such misapprehensions decreases support for content moderation. The Article concludes with brief remarks on what policy guidance can be drawn from the experimental results.

---

24. Alizadeh et al., *supra* note 4, at 6 ("To rise to the political agenda, a given issue must first be construed as politically salient and specific arguments put forward as to how and why it might warrant policy intervention. Therefore, how political actors frame content moderation may impact the kinds of solutions proposed. For example, if content moderation is primarily framed as a violation of free speech, policymakers might be more hesitant to implement strict regulation on platforms' rules around hate speech, misinformation and sensitive content.").

## I. A RIGHTS-BASED NARRATIVE FROM AND ABOUT PLATFORMS

From a historical perspective, it is understandable that laypeople could have acquired inaccurately broad notions regarding the scope of applicability of their constitutional rights vis-a-vis digital platforms. The internet revolution ran on a rights-based narrative for decades, including into the birth of the current platform economy. As documented by Jonathan Zittrain[25] and Evelyn Douek,[26] digital content platforms—and internet companies more generally—relied for quite some time on vigorous free speech philosophies and a lack of speech regulatory practices, with such an approach deeply grounded in rights-based notions applied to online speech. Digital business ventures saw and described themselves as champions of free expression. This rights-based language and correlating minimalist practice of online speech governance were grounded in distinctly American notions of free speech, which are in turn strongly tied to the First Amendment.[27]

Evelyn Douek has recently argued that content platforms' self-image as protectors of speech rights, and the consequential minimal speech regulation they undertook in the past, has recently evolved into an approach that has moved away from absolutist, rights-based notions of free expression, and towards notions of proportionality and probability.[28] Douek consequently encourages policymakers to understand the mechanics of this new reality so that they can create policy reflecting the business and technical realities of content moderation as currently

---

25. Jonathan Zittrain, Three Eras of Digital Governance (Sept. 15, 2019) (unpublished manuscript) (on file with SSRN), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458435. Zittrain describes "The Rights Era" as quite relevant to speech: "The original consideration of threats as external to the otherwise-mostly-beneficial uses of tech made for a ready framing of Internet governance issues around rights, and in particular a classic libertarian ethos of the preservation of rapidly-growing individual affordances in speech—'now anyone can speak without a gatekeeper!'—against encroachment by government censorship or corporate pushback motivated by the disruption of established business models." *Id.* at 1–2 (internal citations omitted).

26. Douek, *supra* note 13.

27. *Id.* at 771.

28. *Id.* at 766.

implemented. Douek is right; policymaking should reflect practical facts on the ground, as they are right now.

But Douek does not discuss whether ordinary Americans have moved on from a rights-based notion of online expression. Recent empirical evidence suggests that they have not.[29] One motivating concern of this Article is that the beliefs and opinions of laypeople are underexplored in the scholarly debate about the future of online speech regulation, despite the clear fact that platforms and politicians alike should account for public opinion when crafting policy.

For two reasons, it is important to understand the extent to which the general public sees expression on digital platforms as a matter of free speech rights. First, it is possible that laypeople have not moved on from a rights-based view of online expression even if platforms have. This would suggest an important source of dissonance between platform policies and platform customers—one that could lead to more conflict than cooperation if platform users fundamentally disagree with the operating paradigm a platform adopts for online speech governance.

Second, rights-based language is important and powerful regardless of whether it creates a conflict between platforms and their users. Rights-based language is important to ordinary Americans[30] and can be particularly influential as to what is prioritized in policymaking.[31] Thus, if we find that rights-based thinking is important to laypeople in their assessment of online speech governance, we can imagine this will shape how laypeople respond to existing content moderation policies as well as to future policy developments.

---

29. Alizadeh et al., *supra* note 4.

30. "However articulated, defended, or accounted for, the sense of legal rights as claims whose realization has intrinsic value can fairly be called rampant in our culture and traditions." Elizabeth M. Schneider, *The Dialectic of Rights and Politics: Perspectives from the Women's Movement*, 61 N.Y.U. L. REV. 589, 589–90 n.1 (1986) (citing Frank I. Michelman, *The Supreme Court and Litigation Access Fees: The Right to Protect One's Own Rights - Part I*, 1973 DUKE L.J. 1153, 1177 (1974), https://www.jstor.org/stable/1371708).

31. For example, in an influential article, Elizabeth M. Schneider discussed her experiences with rights-based language in the feminist movement, highlighting "the rich, complex, and dynamic process through which political experience can shape the articulation of a right, and the way in which this articulation then shapes the development of the political process." *Id.* at 590.

### *A. Surveys on the First Amendment and Content Moderation*

There is a longstanding tradition of using surveys to assess lay beliefs and opinions about the First Amendment. The results of these surveys are largely consistent in demonstrating contradictory impulses in the American public in recent years. Many survey respondents agree both with the idea that digital platforms have an obligation to address problematic content, but many Americans also agree with the notion that content moderation infringes on free expression.[32] Further, survey responses consistently suggest that many Americans have inaccurate beliefs with respect to the application of First Amendment restrictions.

Consider a 2019 YouGov poll of a representative sample of 1,245 Americans. More than half of Republican respondents and about one third of Democratic respondents indicated that removing content or comments from a social media platform suppresses free speech.[33] Yet a high portion of respondents also believed that social media companies should monitor user-created content and comments, with 45 percent of survey respondents agreeing that social media companies "have a responsibility to protect the public from objectionable content."[34] While the YouGov survey did not ask respondents explicitly about the First Amendment, these results nonetheless highlight strong but contradictory impulses from the American public regarding online speech, towards both more and less regulation of online speech by platform companies.

A 2020 survey of 3,000 Americans conducted by the Freedom Forum[35] found similarly conflicting desires from the public. Sixty-nine percent of respondents agreed that social media platforms should be responsible for

---

32. Unfortunately, existing survey work does not report on the degree of agreement on a within-subjects basis with these conflicting imperatives.

33. Jamie Ballard, *Most Conservatives Believe Removing Content and Comments on Social Media Is Suppressing Free Speech*, YOUGOVAMERICA (Apr. 29, 2019, 7:45 AM), https://today.yougov.com/topics/technology/articles-reports/2019/04/29/content-moderation-social-media-free-speech-poll. The survey did not clearly define "free speech" or explicitly tie it to the First Amendment.

34. *Id.*

35. The Freedom Forum is a non-partisan organization that has organized a yearly First Amendment survey for decades. *What We Do*, FREEDOM F., https://www.freedomforum.org/what-we-do/ (last visited Apr. 4 2022).

content,[36] confirming the strong impulse towards platform responsibility for content evidenced in the 2019 YouGov survey. Yet, 70 percent of survey respondents considered "Big tech companies" to represent at least a small threat to the First Amendment. 23 percent of respondents classified such companies as "a significant threat," the highest level of threat.[37] Thus Americans saw platforms as simultaneously responsible for policing content but also as threatening free expression. Such positions are not necessarily logically inconsistent, but they highlight the difficult position platforms find themselves in if they seek to please Americans.

The Freedom Forum survey also gauged knowledge and attitudes about the First Amendment. Only 36 percent of participants correctly indicated that platforms have their own First Amendment rights related to practices of content moderation.[38] The Freedom Forum survey also identified two widespread misconceptions related to the applicability of the First Amendment to private actors: (1) that First Amendment protections apply in private workplaces (74 percent incorrectly believed this); and (2) that First Amendment protections prevent the firing of athletes for their political speech (66 percent incorrectly believed this).[39] The authors of the Freedom Forum survey report speculated that such incorrect beliefs could be due in part to "recent high-profile examples in which athletes incorrectly raised First Amendment arguments when they were fired, disciplined or their contracts were not renewed based on their political activity."[40]

Such speculation demonstrates the plausibility of influential politicians creating or exploiting constitutional misunderstandings through their false claims of First Amendment infringements by platforms. If it is reasonable

---

36. FREEDOM F., THE FIRST AMENDMENT: WHERE AMERICA STANDS 10 (2020), https://survey.freedomforum.org/content/uploads/2021/09/Freedom-Forum-Downloadable-Report.pdf.

37. *Id.* at 24.

38. *Id.* at 16.

39. *Id.* at 12 ("Americans have a good understanding of how the First Amendment protects from overt government interference but are less sure when it comes to questions of how the First Amendment can or cannot limit the actions of businesses, schools or individuals.").

40. *Id.* at 13 ("This misconception may derive from recent high-profile examples in which athletes incorrectly raised First Amendment arguments when they were fired, disciplined or their contracts were not renewed based on their political activity. Only 34% correctly said professional athletes, who are private employees, could be fired.").

to speculate that prominent athletes can influence beliefs about the scope of constitutional protections, it likewise seems plausible that prominent politicians (and their lawsuits or tweets) can create incorrect beliefs in laypeople regarding the breadth of constitutional protections. Thus, this report also highlights the ways in which politicians' constitutionalization of platform moderation policy debates could be influencing the beliefs and opinions of laypeople.

Another recent survey likewise found that laypeople misunderstand the scope of constitutional speech protections. The Knight Foundation commissioned a nationally representative survey of 4,000 Americans that took place in the summer of 2021—that is, months after the January 2021 Capital Riot. As described earlier, that event seemed to catalyze significant discussion of content moderation, particularly among laypeople. The event also catalyzed significant (accurate) coverage by the mainstream press discussing the application of First Amendment restrictions to digital platforms. It is conceivable that after such a watershed moment public attitudes or constitutional literacy levels could have changed. Nonetheless, the results of the Knight Foundation survey were broadly consistent with both the 2019 YouGov survey and the 2020 Freedom Foundation survey.

The recent Knight Foundation survey offers yet more evidence of First Amendment misapprehensions. Consider that, in the case of five questions relating to whether the First Amendment barred private actors from restricting speech, large minorities answered the question incorrectly in every case.[41] Two of these questions directly related to the applicability of the First Amendment to social media companies: the first question, "[b]arring someone from social media is a violation of their First Amendment rights," was answered incorrectly by 35 percent of respondents; the second question, "[t]he First Amendment prevents social media companies (such as Facebook, Instagram, TikTok, Twitter, YouTube) from punishing someone for making offensive statements on their platforms," was answered incorrectly by 28 percent of respondents.[42]

---

41. Among the five questions, incorrectness rates ranged from 18 to 40 percent of participants. KNIGHT FOUND. – IPSOS, FREE EXPRESSION IN AMERICA POST-2020: A LANDMARK SURVEY OF AMERICANS' VIEWS ON SPEECH RIGHTS 17 (2021), https://knightfoundation.org/wp-content/uploads/2022/01/KF_Free_Expression_2022.pdf. Two of the questions applied directly to the case of digital platforms. *Id.*

42. *Id.* (Each question was presented in a true/false format. The correct answer to both questions, of course, was "false").

These post-January 6 responses provide further and more specific evidence to motivate the study presented in this Article.

The Knight Foundation survey also identified partisan disparities consistent with those reported in the YouGov survey. Thus, despite the volume of discourse on social media (which could in theory have facilitated the development of a more unified consensus on the topic) and despite efforts at constitutional education undertaken by news organizations in the wake of the January 6 2021 Capitol attack (which could in theory have significantly increased First Amendment literacy), the same misperceptions and partisan disparities persisted.

Like the YouGov survey, the Knight Foundation survey explored partisan differences in beliefs about First Amendment protections, and identified wide differences between Democrats and Republicans when assessing whether certain behaviors constitute legitimate examples of expression protected by the First Amendment. For example, 57 percent of Republicans, but only 20 percent of Democrats, believed that "[p]eople spreading misinformation about the 2020 election results online" were engaging in protected expression,[43] a 37 percentage point partisan gap. There was likewise a 24 percentage point gap between Democrats and Republicans regarding the question of whether spreading COVID-19 misinformation online constituted a legitimate example of someone expressing their First Amendment Rights. Both the YouGov poll and Knight Foundation poll suggested that Republicans had broader notions than Democrats regarding the scope of constitutional speech protections.[44]

It is also possible to study American attitudes by reviewing social media content. Twitter's permanent suspension of Donald Trump significantly increased the public's interest in online speech governance, as documented by Meysam Alizadeh in a recent working paper. Alizadeh studied public tweets about content moderation from January 2020 through April 2021. Alizadeh found that the January 6th Capitol Attack led to the

---

43. *Id.* at 21.

44. It is possible that the particular questions demonstrated areas where Republicans were more likely to express a belief in broad protections and that other questions would have shown the opposite effect. However, the topics chosen by the survey seem motivated by topicality driven by current events, and likely did not intend to produce this observed partisan effect.

largest volume of English-language discussion about the topic on Twitter.[45] Non-expert users generated the bulk of discussion on content moderation, further showing that the issue was one of general interest, not limited to commentary by experts, politicians, or journalists.[46] In fact, the proportion of non-experts discussing the topic jumped significantly following Twitter's suspension of Trump's account, suggesting that the event generated greater new interest among non-experts, showing its cultural importance[47]

Because the event created an expanded dialogue (at least in online venues) regarding content moderation, Trump's account suspension also created a substantial body of textual data to study how laypeople talk about the topic. Notably, discussions were quite different in different Twitter communities: Conservatives' most popular hashtags in relevant tweets related to the Second Amendment,[48] BigTech, and Section 230; liberals emphasized "the Big Lie," Joe Biden, and Section 230, as well as some less directly relevant hashtags, such as #DiaperDon and #transgender.[49]

---

45. Alizadeh et al., *supra* note 4, at 9. The authors undertook a comprehensive analysis of all discussion of content moderation on Twitter, as identified by Twitter's academic API. *Id.* at 2. They found, "[t]he salience of content moderation was very low until Twitter started fact-checking Donald Trump's tweets in June 2020. After that, the issue only received a moderate amount of attention, as can be seen in the number of retweets. The peak was reached, unsurprisingly, when Donald Trump was initially banned from Twitter for incitement of violence in the context of the assault on the Capitol on January 6, 2021." *Id.*

46. *Id.* at 10.

47. *Id.*

48. References to the Second Amendment do not seem likely to be directly related to content moderation but could reflect a general interest or basis for affinity of the conservative communities surveyed on Twitter. I take the Second Amendment hashtag's popularity in discourse on content moderation as evidence of the importance of constitutional issues to the community generally.

49. Alizadeh et al. identified community clusters within the discourse on content moderation. Alizadeh et al., *supra* note 4, at 11–13. The most popular hashtags in the liberal leaning cluster were "#DiaperDon (referring to unsubstantiated claims about Trump using adult diapers), #BigLie, #Section230, #transgender, and #JoeBiden." *Id.* at 12. The most popular hashtags in the conservative leaning cluster were "#Section230, #BigTech, #antifa, #Twitter, and #2A (referring to the Second Amendment)." *Id.* Among

Alizadeh's work provides a quantitative basis for the notion that laypeople's[50] conceptual understanding of online speech regulation varies significantly with partisan identification.[51]

In summary, the available empirical research establishes a few key points of motivation for this Article. First, many Americans inaccurately believe that First Amendment restrictions extend to digital platforms. Second, and in contradiction, many Americans believe that digital platforms have an obligation to police content. Finally, there are significant partisan disparities in judgments about online speech and in the concepts laypeople find most relevant to the question of online speech regulation.

This constellation of attitudes suggests an impossible situation for content platforms. Could it be that the false beliefs about the First Amendment are themselves a source of trouble in the debate? Perhaps

---

the conservative-leaning community #2A was also one of the most frequent tags in the user profile, which was not the case in the liberal-leaning community. *Id.* at 12.

50. It is difficult to know to what extent Twitter users are representative of laypeople. Existing research suggests that Twitter users are systematically different from the population as a whole in that they are younger, more educated, and more interested in politics than the general population. Jonathan Mellon & Christopher Prosser, *Twitter and Facebook Are Not Representative of the General Population: Political Attitudes and Demographics of British Social Media Users*, 4 RSCH. & POL. 1, 2–3 (2017), https://journals.sagepub.com/doi/pdf/10.1177/2053168017720008. However, when controlling for age, political affiliation, and education level, there does not appear to be a difference between social media users and non-users regarding voting behavior or political values, at least in a study conducted in the U.K. *Id.* at 3. It has been recently estimated that roughly 20 percent of Americans use Twitter, suggesting that Twitter users constitute an important sample of individuals even if they are not representative of all Americans. David Lazer et al., *Meaningful Measures of Human Society in the Twenty-First Century*, 595 NATURE 189, 192 (2021), https://doi.org/10.1038/s41586-021-03660-7 (internal citation omitted). Researchers have also noted that, while Twitter's users may not be representative of the overall U.S. population, "this population has an outsized influence on the trajectory of public discussion—particularly as the media itself has come to rely upon Twitter as a source of news and a window into public opinion." Christopher A. Bail et al., *Exposure to Opposing Views on Social Media Can Increase Political Polarization*, 115 PNAS 9216, 9220 (2018), https://www.pnas.org/content/115/37/9216.

51. Specifically, the hashtags of #2A for the Second Amendment and #Section230 for Section 230 of the Communications Decency Act were both commonly used in conservative communities on Twitter, while #Section230 was also commonly used by liberal communities on Twitter when discussing content moderation. Alizadeh et al., *supra* note 4, at 12.

some animus against content moderation results from an incorrect perception that content moderation is unconstitutional, rather than from more fundamental objections to content moderation.

### *B. Legal Education: A Possible Move Forward?*

The widespread misunderstanding as to the scope of First Amendment restrictions suggests that platforms might benefit by explaining the legality of their actions to laypeople. Perhaps many Americans would even be relieved to find that they can embrace content moderation by private firms without trampling on the Constitution. The apparent lack of First Amendment literacy as it applies to a highly contentious and socially important debate suggests that educational efforts could be an effective and long overdue response to the contentious policy debate.

As a general concept, First Amendment education is certainly not a novel proposal.[52] Efforts to educate high school students about their First

---

52. There are many First Amendment literacy efforts geared towards lay people. For an example of high school training developed by an advocacy organization, see *The First Amendment in Public Schools*, ANTI-DEFAMATION LEAGUE, https://www.adl.org/education/educator-resources/lesson-plans/the-first-amendment-in-public-schools (last visited Apr. 4 2022), the Anti-Defamation League's First Amendment Training materials for high school students, which includes four lesson plans. The New York Times also partnered with the National Constitution Center to create a lesson plan about the First Amendment. Staci Garber, *Freedom of Speech? A Lesson on Understanding the Protections and Limits of the First Amendment*, N.Y. TIMES (Sept. 12, 2018), https://www.nytimes.com/2018/09/12/learning/lesson-plans/freedom-of-speech-a-lesson-on-understanding-the-protections-and-limits-of-the-first-amendment.html. Likewise, USA Today ran an editorial urging the need for more civic education to correct American misunderstandings of the First Amendment:

> Without a greater emphasis on civic education, and First Amendment rights in particular, many of us will continue to lack the knowledge and tools we need to fully participate in our governance, and taxpayers will continue to foot the bill for legal challenges to state laws that are plainly unconstitutional – laws that should never have been proposed or passed in the first place.

Amy Kristin Sanders, Opinion, *We Need to Do a Better Job Teaching Citizens About the First Amendment*, USA TODAY (Oct. 22, 2021), https://www.usatoday.com/story/opinion/2021/10/22/free-speech-week-first-amendment/8475177002/.

Amendment rights have existed for decades, and in diverse forms.[53] More recently, at least one scholar has proposed First Amendment literacy as an antidote to the proliferation of obviously unconstitutional laws targeted at digital platforms in recent years.[54] In a 2021 editorial in *USA Today*, Amy Kristin Sanders argued in favor of greater civic education as an important tool to reduce the recent practice of politicians enacting clearly unconstitutional laws.

Without a greater emphasis on civic education, and First Amendment rights in particular, many of us will continue to lack the knowledge and tools we need to fully participate in our governance, and taxpayers will continue to foot the bill for legal challenges to state laws that are plainly unconstitutional—laws that should never have been proposed or passed in the first place.[55]

Sanders identified a specific price American society pays due to a lack of familiarity with First Amendment black letter law: a political environment in which politicians are not penalized —and may even be rewarded—for implicitly propagating constitutional misinformation.

This Article examines the possibility that Americans may object to content moderation due to a misunderstanding regarding the scope of First Amendment restrictions. Just as Sanders is concerned that Americans do not discipline their politicians enough due to a lack of First Amendment knowledge, I explore the concern that Americans may support (or oppose) content moderation less (more) than they otherwise would if they had correct information about the reach of First Amendment restrictions.[56]

---

53. Even a U.S. federal court has created educational materials for high school students to learn about their First Amendment rights. *See generally* U.S. DIST. CT., DIST. OF MONT., FIRST AMENDMENT IN SCHOOLS, https://www.mtd.uscourts.gov/sites/mtd/files/FirstAmendCluster.TeachInst3Cases.pdf. For a listing of currently available First Amendment trainings see, Brian J. Buchanan, *The 12 Best Sites for Teaching the First Amendment*, FREE SPEECH CTR. (Jan. 13 2020), https://mtsu.edu/first-amendment/post/457/the-12-best-sites-for-teaching-the-first-amendment.

54. *See* Sanders, *supra* note 52 (citing examples of clearly unconstitutional laws recently passed in Florida, Arizona, and Texas).

55. *Id.*

56. This experiment thus also touches on the expressive power of law. When participants find that a particular kind of law—in this case, constitutional law—does not forbid or

Perhaps if ordinary Americans understood the limitations of constitutional restrictions, they would be less likely to believe that such practices threaten free speech and therefore less opposed to content moderation.

Education appears to be an underexplored option as a means to enhance public acceptance of content moderation policies. This is particularly surprising given the strong connection between the First Amendment and the widely held tenets of liberal democratic societies that more information[57] and more education[58] are inherently desirable, as embodied in notions such as the marketplace of ideas.

What's more, education has long been understood as having instrumental value as well as being an intrinsic good. For example, in the 1950s, education efforts undertaken by local volunteers were used to

---

otherwise implicitly condemn a practice, does this make the practice itself more acceptable to them? *Cf.* Maggie Wittlin, Note, *Buckling Under Pressure: An Empirical Test of the Expressive Effects of Law*, 28 YALE J. ON REGUL. 419 (2011), https://openyls.law.yale.edu/handle/20.500.13051/8138 (finding an effect of the expressive power of law distinct from the law's jurisdictional range and distinct from personal interactions with law enforcement with respect to that law). However, given that the current experiment revolves around what the law is *not*—specifically that the Constitution does not limit private actors attempting to control speech—the experiment can be seen as a test of the negative power of expression of the law as possibly influencing laypeople's own moral or policy judgments.

57. *See* Stanley Ingber, *The Marketplace of Ideas: A Legitimizing Myth*, 1 DUKE L.J. 1, 2 nn.1–2 (1984), https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=2867&context=dlj (references in n.1 and n.2 contain an extensive catalog of academic and judicial discussions of this notion). Note that Ingber concludes that the theoretical underpinnings of this argument are flawed, but nonetheless documents the importance and wide acceptance of this notion. *Id.* at 4–5.

58. The widespread belief in modern democratic societies in the importance and benefits of education is most notably evinced in mandatory universal education and significant government subsidies even for advanced education. For more than a century, American educational theorists have also argued the importance of education generally, and civic education in particular, to the proper functioning of society. For example, Horace Mann, wrote in 1848 in support of education and its attendant civic benefits, "It may be an easy thing to make a Republic; but it is a very laborious thing to make Republicans; and woe to the republic that rests upon no better foundations than ignorance, selfishness, and passion." Rebecca Winthrop, *The Need for Civic Education in 21st-Century Schools*, BROOKINGS (June 4, 2020), https://www.brookings.edu/policy2020/bigideas/the-need-for-civic-education-in-21st-century-schools/.

overcome polio vaccination hesitancy.[59] Likewise, educational interventions have been deployed to fight racism[60] and reduce poverty,[61] with at least some success.[62] Of course, educational interventions can sometimes be ineffective or even harmful,[63] and the possibility that education may backfire relative to its policy goal points to the need to test educational interventions rigorously rather than assuming they are inevitably effective.[64]

In recent years, educational interventions have been a common technique in Western nations when addressing threats associated with the digital world. For example, many Western nations, including France, Finland, Sweden, and Denmark, have initiated national efforts to include secondary school educational modules teaching students to identify online

---

59. Susan Brink, *Can't Help Falling in Love with a Vaccine: How Polio Campaign Beat Vaccine Hesitancy*, NPR (May 3, 2021, 9:00 AM), https://www.npr.org/sections/health-shots/2021/05/03/988756973/cant-help-falling-in-love-with-a-vaccine-how-polio-campaign-beat-vaccine-hesitan. ("The polio vaccine effort offers some lessons for today,' says Stewart. First, volunteers from local communities are trusted and invaluable in providing education on disease, research and vaccines.").

60. *See, e.g.*, Elizabeth Vera et al., *Education Interventions for Reducing Racism*, *in* THE COST OF RACISM FOR PEOPLE OF COLOR: CONTEXTUALIZING EXPERIENCES OF DISCRIMINATION 295–316 (A.N. Alvarez et al. eds., 2016).

61. *Eradication of Global Poverty Begins with Education*, NYU DISPATCH (May 22, 2018), https://wp.nyu.edu/dispatch/2018/05/22/eradication-of-global-poverty-begins-with-education/.

62. For a review of some evidence that education reduces racism, see John Duckitt, *Reducing Prejudice: An Historical and Multi-Level Approach, in* UNDERSTANDING PREJUDICE, RACISM AND SOCIAL CONFLICT 253 (Martha Augoustinos & Katherine J. Reynolds, eds., 2001).

63. For example, it is commonly asserted that employee trainings to reduce sexual harassment are ineffective or possibly harmful. *See, e.g.*, Frank Dobbin & Alexandra Kalev, *Why Sexual Harassment Programs Backfire: And What to Do About It*, HARV. BUS. REV., May–June 2020, at 45, https://hbr.org/2020/05/why-sexual-harassment-programs-backfire.

64. I was not able to identify any First Amendment training materials that were rigorously tested to determine whether they met policy objectives. However, this is not particularly surprising given that randomized controlled testing is not a customary practice for educational materials generally.

misinformation, with an emphasis on teaching critical thinking skills.[65] Likewise, private firms and consumer protection regulators have looked to educational interventions when seeking to enhance cybersecurity.[66] Given that educational interventions are already widely used to address digital threats, a simple educational intervention could be a reasonable initial policy solution to the speech regulation bind technology companies currently face.

## II. EXPERIMENTAL STRATEGY

Laypeople have conflicting desires. They want platforms to assure some minimal level of online discourse, but they seem to worry that platforms infringe the Constitution by taking direct action regarding problematic speech. This Article tests a series of hypotheses[67] inspired by these empirical observations:

- Hypothesis One (H1): Participants who express correct beliefs about the scope of First Amendment restrictions will express higher support for a platform's content moderation action than those who express incorrect beliefs (the Constitutional Content Moderation Connection hypothesis).

---

65. The French government has been allocating an educational budget for teaching students to identify misinformation online since 2015. *See* Adam Satariano & Elian Peltier, *In France, School Lessons Ask: Which Twitter Post Should You Trust?*, N.Y. TIMES (Dec. 13, 2018), https://www.nytimes.com/2018/12/13/technology/france-internet-literacy-school.html. Likewise, Finland, Sweden, and Denmark have included curriculum elements to teach "critical thinking about misinformation to schoolchildren" since at least 2019. *See* Emma Charlton, *How Finland is Fighting Fake News - in the Classroom*, WORLD ECON. F. (May 21, 2019), https://www.weforum.org/agenda/2019/05/how-finland-is-fighting-fake-news-in-the-classroom/.

66. *See, e.g.*, *Cybersecurity Basics*, FED. TRADE COMM'N, https://www.ftc.gov/tips-advice/business-center/small-businesses/cybersecurity/basics (last visited Mar. 11, 2022) ("Create a culture of security by implementing a regular schedule of employee training. Update employees as you find out about new risks and vulnerabilities. If employees don't attend, consider blocking their access to the network.").

67. These hypotheses are not identical to those in the pre-registration but are adapted for simplicity of narrative. The hypotheses as stated, largely map onto the pre-registered hypotheses; Aileen Nielsen, *The Rights and Wrongs of Folk Speech Beliefs and Content Moderation (#84896)*, PENN. WHARTON CREDIBILITY LAB: ASPREDICTED (Jan. 24, 2022, 2:15 PM), https://aspredicted.org/a7iu2.pdf [hereinafter Nielsen, *Q&A*].

- Hypothesis Two (H2): Many participants will express inaccurately broad beliefs[68] about the scope of First Amendment restrictions, but these incorrect beliefs can be corrected with targeted information (the Correctable Belief hypothesis).
- Hypothesis Three (H3): Exposure to Constitutional Information regarding the lack of First Amendment restrictions on private entities will increase support for a platform's content moderation action (the Connection Manipulation hypothesis).

These hypotheses were tested by an online vignette Experiment.[69] The full design and flow of the experiment are presented in Figure 1. For conceptual clarity, the experiment can be understood as having four stages (though the participants experienced them as a continuous flow of screens in a single online interface):

- Stage 1 was the point in the experiment at which constitutional knowledge was tested or constitutional information was provided.[70]
- Stage 2 was the point in the experiment in which all participants read a vignette about a platform's decision to suspend a user account after the user repeatedly violated the platform's rules. Participants also provided judgments about the parties' actions.[71]
- Stage 3 elicited descriptive or normative beliefs about the scope of constitutional protections, as a robustness check.[72]

---

68. Specifically: that First Amendment restrictions apply to private entities.

69. A vignette experiment is one in which participants read about a hypothetical situation and contemplate their reactions to that situation. Although vignette studies cannot replicate all the salience and complexities of the real world, they can be quite informative for predicting real world behavior. Vignettes can be particularly useful for abstracting away from problematic real-world details to measure general behaviors and beliefs. This is helpful in the current political climate because most real-world examples of content moderation decisions take place in highly politicized contexts.

70. *See* Aileen Nielsen, *Supplementary Materials*, OSF 13–16, https://osf.io/52qk4/ (last visited Mar. 4, 2022) [hereinafter Nielsen, *Supplements*].

71. *See id.* at 16–17.

72. *See id.* at 17–18.

- Stage 4 surveyed demographic and political affiliation information, after all experimental metrics of interest had been collected.[73]

The experiment implemented a 3 x 2 factorial design. The three-way factor, implemented in Stage 1 of the experiment (see Figure 1) was a manipulation of exposure to constitutional information. Before proceeding to the vignette about a content moderation decision, participants were assigned to Constitutional Information, Knowledge Elicitation, or the Control Group. The two-way factor, implemented in Stage 3, was a robustness check after participants responded to the content moderation vignette, eliciting opinions regarding the scope of constitutional speech protections either as a normative or as a descriptive matter.

In Stage 1, participants were randomly assigned with equal probability to one of three treatments: (1) a short constitutional law training (the Constitutional Information treatment); (2) a question about constitutional law (the Knowledge Elicitation treatment); or (3) no information or questions regarding constitutional protections (the Control Group treatment). Stage 1 provided the opportunity to test H2 (the correctable belief hypothesis) by comparing the rate of correct responses in the Constitutional Information condition as compared to the Knowledge Elicitation condition. If H2 holds, the correctness rate should be higher in the Constitutional Information condition.

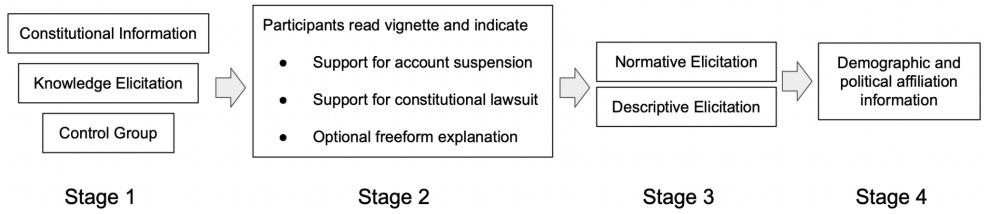



**Figure 1: Experiment design. Vertically stacked boxes indicate between-subject manipulations.**

73. *See id.* at 18. Collecting demographic information at the end of the experiment ensures that elements of identity are not made salient at the time that the key variables of interest are measured, since the opposite sequence could create a demand effect. In this case, priming participants regarding their demographic or political identity could artificially introduce salience of these categories and prompt participants to possibly make inferences regarding their expected behavior, which in turn could create apparent effects of demographic or partisan information. *See generally* Daniel John Zizzo, *Experimenter Demand Effects in Economic Experiments*, 13 EXPERIMENTAL ECON. 75 (2010), https://link.springer.com/article/10.1007/s10683-009-9230-z (explaining the concept and when demand effects constitute a challenge to experimental findings).

Participants in the Constitutional Information condition read a short explanation, spaced over multiple screens, describing how constitutional free expression protections do not apply against private organizations. To test comprehension, participants answered a multiple-choice question on whether a private university infringed on students' constitutionally protected freedom of expression by refusing to allow them to invite a controversial speaker to campus. Participants were provided with feedback as to whether they had answered correctly.

Participants in the Knowledge Elicitation condition did not read any information but responded to the same question about the private university. Participants in the Knowledge Elicitation condition were not given feedback regarding the correctness of their answers. Control Group participants did not encounter any information about First Amendment protections or any questions about constitutional protections. Table 1 provides a brief description of the three treatments and labels that will be used to refer to participants, depending upon which treatment they received.

| **Treatment Name** | **Description** |
|---|---|
| Control Group | No information or questions. |
| Knowledge Elicitation | A constitutional law question. |
| Constitutional Information | Information about limited reach of constitutional speech protections, a constitutional law question, and feedback on correctness of question response |

**Table 1: Stage 1 treatments**

In the course of the analysis presented below, participants in the Constitutional Information are sometimes compared to participants in the Knowledge Elicitation condition and sometimes to participants in the Control Group. Comparison of Constitutional Information with the Control Group is preferred because this comparison best models the difference between those who received the information training and those who did not have any prior prompting about constitutional issues. Comparing the Constitutional Information participants to the Control Group is the better available predictor of external validity. However, when it is necessary to

know a participant's constitutional knowledge, it is necessary to use the Knowledge Elicitation condition rather than the Control Group.[74]

In Stage 2, all participants read the same content moderation scenario.[75] The scenario described a popular user on a news-crowdsourcing digital platform. After several warnings, the user is permanently suspended from the platform due to multiple instances of posting false information in violation of the platform's terms of service. The nature of the hypothetical platform and the hypothetical user's identity and infractions were kept vague so as to encourage participants to judge the facts on their own merit rather than with reference to prior real-world events.

After reading the vignette, participants were asked whether they supported the decision to remove the user and whether they supported a constitutional lawsuit by the suspended user against the platform.[76] Responses were provided on a 5-point Likert scale, ranging from Strongly Disagree to Strongly Agree.[77] Participants were also given the option to provide a freeform response explaining their choice.

In combination with Stage 1, Stage 2 provided the opportunity to test H1, the Constitutional Content Moderation Connection hypothesis, and H3, the Connection Manipulation hypothesis. H1 could be tested by

---

74. A possible critique of this is that the measure of constitutional knowledge could be taken in Stage 3, after reading the content moderation scenario, rather than in Stage 1. This possibility was rejected. Such an alternative sequencing would make the Knowledge Elicitation group less comparable to the Constitutional Information group due to the possibility that ordering effects would be different in a group that contemplated content moderation before a constitutional question as compared to one that contemplated that same scenario after a constitutional question.

75. The full text is presented *infra* Section II.A.b.

76. The order of these two questions was counterbalanced so that participants were equally likely to see either sequence of the two questions.

77. A Likert scale is a standard experimental tool used to measure participant attitudes directly by allowing participants to express the strength of their agreement or disagreement with a particular idea or statement. These are then mapped to numbers, which are presumed to have a rank order. While Likert scale data need not necessarily reflect equal intervals between different adjacent points on the scale, they are commonly treated as such for purposes of analysis (such as computing means or standard deviations). *See* Huiping Wu & Shing-On Leung, *Can Likert Scales be Treated as Interval Scales? – A Simulation Study*, 43 J. SOC. SERVS. RSCH. 527, 528 (2017).

comparing the level of support for content moderation in the Knowledge Elicitation condition among those who answered the constitutional law question correctly and those who did not. H3 could be tested by comparing the level of support for content moderation of those in the Constitutional Information condition to those in the Control Group condition.

Stage 3 posed two questions[78] to participants regarding their beliefs about constitutionally protected freedom in general, and social media in particular. The questions were framed either to elicit descriptive beliefs about the state of the law or normative beliefs about the desirable state of the law, with participants randomly assigned to only one set of questions. Stage 3 sought to establish whether the effect of the Constitutional Information was sufficiently durable to be detected later in the experiment and, if so, the specificity with which the Constitutional Information modified descriptive rather than normative beliefs about constitutional speech protection.

Finally, in Stage 4, participants provided information about their age, race, gender, and political affiliation before formally exiting the experiment to receive compensation.[79] This information was collected to enable statistical analysis regarding the importance of these demographic and political variables in predicting attitudes. The full experimental text and flow can be found in the online supplementary materials.[80]

---

78. The wording of the questions (counterbalanced in order) was as follows:

My freedom to express myself in all situations [ought to be/is] protected by the U.S. Constitution.

No one [ought to have/has] the legal right to tell me what I can or cannot say on social media.

The normative and descriptive word choices are shown within the square brackets and distinguished by a forward slash.

79. Polling was conducted on the Prolific.co polling platform in January 2022. More information about procedural details and the experiment pre-registration document, as well as the supplementary materials, can be found online. *See* Nielsen, *Q&A, supra* note 67; Nielsen, *Supplements*, *supra* note 70.

80. Nielsen, *Supplements, supra* note 70.

### *A. The Experimental Texts*

There are two key textual components of the experiment. The first is the content provided to those in the Constitutional Information condition of Stage 1. The second is the vignette presented in the online speech scenario of Stage 2. The full text of both are included below, along with commentary.

#### 1. The Constitutional Information

The text of the Constitutional Information was displayed over multiple screens to ensure greater attention by participants. Each bullet below indicates a separate screen, which participants navigate by clicking on a "Next" button.[81]

- In the United States, Constitutional protections of free speech do not apply against private organizations.
- In other words, the Constitution does not prohibit private organizations from limiting speech.
- From a constitutional perspective, private organizations are free to restrict the speech of their members, customers, or employees.
- For this reason, people who face restrictions on speech could not have a constitutional case if a private organization has acted to restrict their speech.

The language may seem unduly simplistic to legal scholars, but it reflects the tenor and content of training materials developed for

81. I have previously found multi-screen structuring of content to be effective for increasing comprehension by participants. This was particularly important in the current experiment, which included a flat rate of payment for participants. I ruled out a performance assessment or other form of attention check due to the ideologically-infused nature of the topic under investigation. It would be likely that incorrect answers (failed attention checks) correlated with relevant viewpoints rather than merely with attention or care in undertaking the experiment. Indeed, the very premise of this experiment is that incorrect answers reveal more than lapsed attention.

laypeople,[82] as well as think tank policy documents.[83] Of course, even situations apparently covered by black letter law can be more complicated than simple binary statements. As discussed in the Introduction, some legal scholars have recently made arguments with respect to the First Amendment possibly restricting content moderation activities by digital platforms under a variety of legal theories.[84] Some might object that the above training does not acknowledge questions about whether a highly regulated, socially powerful entity such as large digital platforms could be restrained by the First Amendment (such as through a close relationship with a federal regulator[85] or by stepping into the traditional role of government in a relevant way[86]). Yet, to date, the Supreme Court has

---

82. *See, e.g.*, *First Amendment Lesson Plan: Free Speech on College Campuses*, MIDDLE TENN. STATE UNIV.: FREE SPEECH CTR., https://www.mtsu.edu/first-amendment/page/free-speech-college-campuses (last visited Mar. 3, 2022). In a scenario that closely mirrors the one presented in the training, regarding students inviting a controversial speaker to campus, the training includes the following bullet points to summarize the lessons:

"Key concepts

1. Public colleges are bound by the First Amendment not to restrict campus speech on the basis of its content.

2. Private colleges are not bound by the First Amendment, but may have policies stating a commitment to free expression on campus." *Id.*

83. *See, e.g.*, John Samples, *Why the Government Should Not Regulate Content Moderation of Social Media*, CATO INST. (Apr. 19, 2019), https://www.cato.org/policy-analysis/why-government-should-not-regulate-content-moderation-social-media ("The First Amendment protects the freedom of speech from state action. Social media are not government and hence are not constrained by the First Amendment. These platforms are protected by the First Amendment but need not apply it to speech by their users.").

84. *See* Volokh, *supra* note 9; Sylvain, *supra* note 9; *Biden v. Knight First Amend. Inst. at Columbia Univ.*, 141 S. Ct. 1220, 1221–27 (2021) (Thomas, J., concurring).

85. *See*, *e.g.*, Pub. Utils. Comm'n v. Pollak, 343 U.S. 451, 461–63 (1952) (finding that such a situation did apply to the highly regulated D.C. streetcar company, but that there was no Constitutional infringement under the facts of the case).

86. *See, e.g.*, Marsh v. Alabama, 326 U.S. 501 (1946) (finding a First Amendment restriction binding a company in the case of a company town because of specific traditional roles of government assumed by the private entity). Yet while Philip Hamburger recently cited Marsh in a 2021 Wall Street Journal opinion piece for the proposition that "[t]he First Amendment protects Americans even in privately owned public forums, such as company towns," Philip Hamburger, *The Constitution Can Crack Section 230*, WALL ST. J. (Jan. 29, 2021, 2:00 PM), https://www.wsj.com/articles/the-constitution-can-crack-section-

consistently declined to expand the extent of First Amendment restrictions to private companies,[87] and groups that develop training materials have found it reasonable to deliver a message in the simple binary logic reflected in the experimental text.[88]

In addition to the Constitutional Information text quoted above, participants encountered a constitutional law question:

- A private university refuses to allow students to invite a controversial speaker to campus. The university leadership fears that hosting this person will lead to a tense environment on campus, and possibly reduce alumni donations.
- Could the university's refusal to allow the speaker potentially be a violation of the students' constitutional right to free speech?

Participants were given the option to answer with the following responses: "Yes," "Maybe,"[89] or "No," with the correct answer of "No"

---

230-11611946851, Berin Szóka and Ari Cohn remind us, "Marsh has been read very narrowly by the Supreme Court, which has declined to extend its holding on multiple occasions and certainly has never applied it to any media company," Berin Szóka & Ari Cohn, *The Wall Street Journal Misreads Section 230 and the First Amendment*, LAWFARE (Feb. 3, 2021, 3:43 PM), https://www.lawfareblog.com/wall-street-journal-misreads-section-230-and-first-amendment.

87. Justice Kavanaugh recently reminded readers that "[M]erely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors subject to First Amendment constraints." Manhattan Cmty. Access Corp. v. Halleck, 139 S.Ct. 1921, 1930 (2019).

88. *See* discussion *supra* Section I.A. (describing answers in binary language of right or wrong when looking to lay understanding of First Amendment applicability).

89. In analyses regarding correctness of beliefs, those who responded with a "Maybe" were included when calculating rates of correctness but were excluded when comparing population means between correctness and incorrectness. The reason for this was as follows. For calculating the rates of correctness, this is defined as the portion of participants who gave the correct answer, and those responding "Maybe" did not provide a correct answer. On the other hand, when comparing participants who answered a question correctly or incorrectly, those who selected "Maybe" were viewed as declining to provide a definite response due to uncertainty. Not surprisingly, the "Maybe" group tended to be intermediate between the "Yes" and "No" respondents on the metrics of interest. In any case, inclusion or exclusion of the "Maybe" respondents did not change the results of evaluating the hypotheses.

coming last. The same question was used in the Knowledge Elicitation condition.

Once Constitutional Information participants finalized their answer, they were told whether their answer was correct. They also saw the following explanatory text:

- The private university's decision could not be a constitutional violation because a private university is a private entity, and so constitutional protections are not applicable.

This question and response represent standard black letter First Amendment law as presented in First Amendment training designed for lay audiences.[90]

2. The Content Moderation Scenario

The full text of the vignette scenario is reproduced below. Again, bullets indicate separate screens, a tactic used to increase the likelihood that participants read and digest the text in full.

- A popular new app, WhatsHappening, has created a new form of social media sharing, whereby users are free to share personal experiences or local news stories of interest. The goal of WhatsHappening is to crowdsource journalistic information. Any user of WhatsHappening who generates interest can earn money for their content creation and/or news reporting.
- One particular user has become a maverick on the new system, with millions of followers. Her leads have frequently been picked up by major journalism outlets. She seems to have ways and means of obtaining very important information before anyone else can break a story.
- Recently this user began posting stories that were later found to be untrue. She was accused of circulating fraudulent information. She was notified several times by WhatsHappening that she was violating their terms of use.

90. *See, e.g.*, *First Amendment Lesson Plan, supra* note 80 (listing as the second key concept of the lesson that "[p]rivate colleges are not bound by the First Amendment, but may have policies stating a commitment to free expression on campus.").

- Finally, after the fourth violation, WhatsHappening suspended[91] this user's account. In response, the user sued WhatsHappening, alleging that her constitutional freedom of speech had been violated.

The participants were asked two questions, counterbalanced in order, about this scenario, with available responses on a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree."

- Do you agree with WhatsHappening's decision to permanently suspend the account?
- Do you agree with the user's decision to sue WhatsHappening for violating her constitutional freedom of speech?

Participants were also given the option to input a freeform response:

- Anything you wish to share about your opinion?

These questions were asked for a number of reasons. First, participants had the opportunity to express support for both or either of the parties directly involved in the conflict between the platform and the user. This treated support for either party on an equal footing. The open-ended text input allowed participants to flag what was important to them or even to raise issues they might have felt were neglected in the experimental line of questioning.

## III. RESULTS AND DISCUSSION

The structure of the experiment was as described in Figure 1, and the full wording and screen flow are available in the online supplementary materials. A representative sample of N = 1003 U.S. adults as stratified by gender, race, and age was collected via the Prolific.co polling platform in January 2022. Participants were on average compensated at a flat rate equivalent to an hourly rate of more than 30% above the U.S. federal minimum wage for completing the experiment. All participants were included in the analysis reported below.

---

91. Of course, there are many ways platforms can moderate content, and the suspension of a user represents a fairly extreme action. The scenario was chosen because participants were expected to have a stronger reaction to a suspension than they would have to a milder form of moderation, such as removal of one post or one comment. Also, the suspension of a user account can be defined more simply than removal or censure of content. Account suspension can be described with a fairly minimal level of detail without the need for distracting details.

The mean age of participants was forty-five, with a standard deviation of sixteen years. Fifty percent of participants identified as female, 48 percent as male, and 1 percent as non-binary, with another 1 percent declining to identify their gender. Seventy-six percent of the population identified as white, 5 percent as Hispanic or Latino,[92] 14 percent as African American, and 7 percent as Asian. Political affiliation was not included in the sampling stratification, but this information was collected at the end of the experiment. Fifty-one percent of respondents identified as Democrats, 17 percent as Republicans, and 29 percent as unaffiliated, with the remaining 3 percent describing themselves as members of third parties.

The experimental procedure and analysis were pre-registered.[93] The analyses presented below correspond to that pre-registration unless flagged as post hoc. Statistical analyses were conducted with the R statistical package.[94] Comparisons of population means were conducted via a Wilcoxon rank sum test. Linear regressions were fitted as least squares regressions. All data and source code are available in the online supplementary materials.[95]

---

92. This portion of self-identified Hispanic or Latino participants is notably lower than that reported in the population as a whole. In the 2020 census, 18.7 percent of respondents self-identified as having a Hispanic or Latino ethnicity. Nicholas Jones et al., *2020 Census Illuminates Racial and Ethnic Composition of the Country*, U.S. CENSUS BUREAU (Aug. 12, 2021), https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html. It is unclear why the polling vendor failed to deliver a representative sample with respect to participants of Hispanic or Latino identity.

93. Pre-registration is a measure social scientists take to increase the replicability of research. Details of the experiment and anticipated results are recorded in advance so that subsequently published results are more likely replicable and less likely to be the result of undertaking multiple analyses until a statistically significant result is identified and then rationalized *ex post*. Post hoc analyses are analyses that were undertaken after seeing the data, and therefore described in the pre-registration.

94. *The R Project for Statistical Computing*, R PROJECT, https://www.r-project.org. (last visited Mar. 20, 2022)

95. Nielsen, *Supplements*, *supra* note 70.

### *A. Correlation of Constitutional Correctness and Support for Content Moderation*

Consider data to address H1, the Constitutional Content Moderation Connection hypothesis. As shown in Table 2, those who expressed an accurate understanding of the scope of constitutional free expression protections in the Knowledge Elicitation condition reported a higher level of support for content moderation[96] than those with an inaccurate understanding of the scope of constitutional speech protections (post hoc, $p < .001$). That is, participants' demonstrated knowledge regarding the state of First Amendment restrictions on private entities correlated with different levels of support for content moderation, thus confirming H1, the Constitutional Content Moderation Connection hypothesis.

| | **Mean Support for Platform's Decision** |
|---|---|
| Correct response | 4.6 |
| Incorrect response | 4.1 |

**Table 2: Mean support for content moderation was higher for those answering the constitutional question correctly.**

This finding demonstrates an empirical connection between correct constitutional knowledge regarding First Amendment restrictions and higher support for content moderation. While it is not possible to establish a causal link between these factors, the association is interesting per se. It is not inevitable that opinions about content moderation be linked to knowledge about the Constitution.[97]

96. For brevity, I use "support for content moderation" to stand in for "support for the content moderation decision depicted in the vignette."

97. It is true that the correct response correlates with the more permissive view of the constitutionality of a content moderation decision. But, in the alternative, participants could believe that something is constitutional and yet a bad idea, possibly even morally repugnant, for other reasons. That was not the case here. Participants not only accepted

In summary, those who expressed accurate knowledge regarding the limits of First Amendment restrictions also expressed[98] higher support for the platform's decision to enforce a speech restriction, empirically validating a connection between beliefs about First Amendment restrictions and support for content moderation.

### *B. Correctness on Constitutional Law Question and Influence of Constitutional Information*

Having established a connection between support for content moderation and correct constitutional question responses, it is next interesting to ask whether more people can be induced to give correct responses on the constitutional question. Consider H2, the Correctable Belief hypothesis. Among participants in the Knowledge Elicitation treatment, 59 percent answered the constitutional question incorrectly. This result was consistent with high rates of incorrectly broad First Amendment beliefs identified in the Freedom Foundation and Knight Foundation survey studies.[99] However, these incorrect responses were correctable at a high rate.

As shown in Figure 2, the Constitutional Information significantly increased the portion of participants correctly answering the constitutional law question ($p < .0001$), confirming H2. Sixty-four percent of Constitutional Information participants answered the constitutional question correctly, while only forty-one percent of those in the Knowledge Elicitation condition did so. So, the participant population did evince an incorrect understanding of First Amendment restrictions, but the Constitutional Information substantially corrected this constitutional misapprehension.

---

the content moderation decision as legal but in fact expressed support for the content moderation decision.

98. Note that accurate knowledge as assessed via correctness on the constitutional question is endogenous. In other words, it is not possible to manipulate whether participants express correct or incorrect beliefs as an experimental treatment.

99. *See supra* Section I.A.

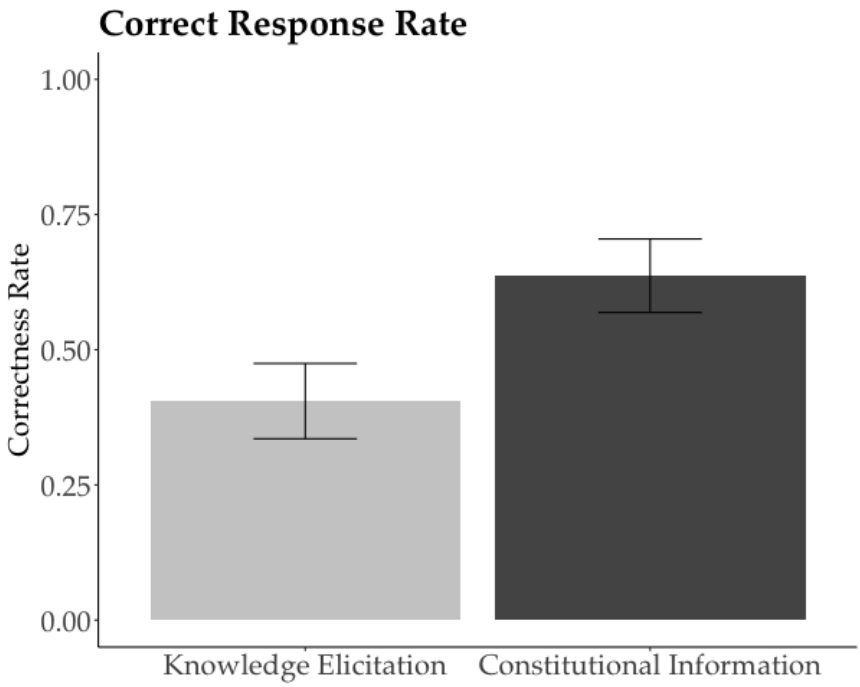



**Figure 2: Constitutional Information increases the rate of correctness for a constitutional law question. Error bars represent +/- standard error.**

An additional question, related to the Correctable Belief hypothesis, is the specificity and durability of the correction induced by the Constitutional Information. Is the effect merely ephemeral and non-specific? One can assess these qualities with the robustness measures collected in Stage 3. Post hoc analysis identified a significant difference between Constitutional Information and Control Group with respect to the descriptive beliefs of the participants (post hoc, $p < .01$) but not with respect to the normative beliefs (post hoc, $p = .8$). In other words, the Constitutional Information affected participants' belief about the state of constitutional protections but not what the state of such protections ought to be. This difference is consistent with the Constitutional Information providing descriptive information but not normative recommendations. Also, the difference in the descriptive beliefs shows that the effects of the Constitutional Information lasted beyond the immediate effects measured in Stage 1. Thus, a robustness check confirms the specificity and durability of the Constitutional Information.

In summary, participants demonstrated inaccurately broad beliefs regarding First Amendment restrictions, but Constitutional Information reduced the rate at which participants demonstrated such beliefs.

### *C. Effect of Constitutional Information on Support for Content Moderation*

So far there are two highly suggestive results. First, untrained participants who expressed accurate beliefs about the scope of First Amendment restrictions are more supportive of a firm's decision to suspend a user for violating online speech rules than those who expressed inaccurate beliefs. Second, inaccurate beliefs about the scope of First Amendment restrictions were corrected in a substantial proportion of participants via the Constitutional Information intervention. This combination of results raises the possibility that educational interventions to correct misapprehensions about the First Amendment's scope, when successful, might also lead to higher support for the platform's decision.[100] I next examine this possibility, expressed in H3, the Connection Manipulation hypothesis.

Surprisingly—and contrary to the pre-registered hypothesis—the Constitutional Information reduced support for the platform's decision. The absolute value of the change was small (from 4.5 to 4.3) but statistically significant ($p < .01$). The downward shift in support manifests as a change in the distribution of expressed support at the highest level of support. The proportion of respondents expressing the highest level of agreement with the suspension of the user's account decreased from 70 percent in the Control Group Condition to 56 percent in the Constitutional Information condition, a statistically significant shift (post hoc, $p < .01$). The Constitutional Information *backfires*, and it apparently backfires among those who otherwise would have been expressing strong support.

This result is consistent with a broad area of literature that looks at motivated reasoning in the presence of unwelcome information, producing backfire effects. An expansive previous literature has examined how people

100. There need not be a direct, causal link from constitutional knowledge to support for content moderation in order for H3 to hold, and I do not make an assertion of a direct causal link. There are also more complicated causal structures consistent with Constitutional Information producing such an outcome.

respond to facts[101] or political opinions[102] contrary to their own beliefs, identifying backfire effects in a wide variety of circumstances. The results here provide proof of such a possibility in response to constitutional education, that is in response to legal facts.

### *D. Political and Demographic Factors*

In previous studies, partisan identity has been a strong predictor of the magnitude and direction of backfire effects.[103] A partisan difference in backfire effects seems likely, too, in the current experiment since inaccurately broad claims of First Amendment violations by private actors

---

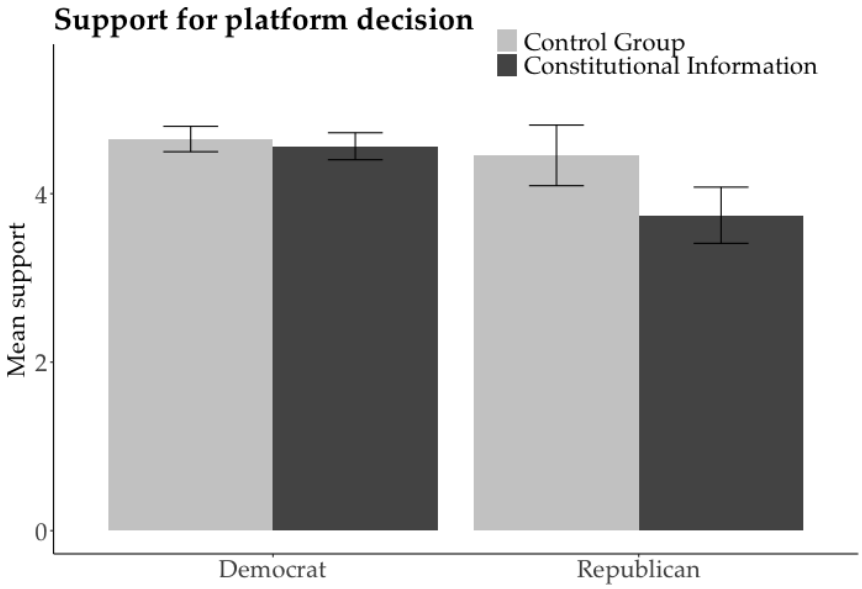
101. For the original description of the backfire effect, *see* Brendan Nyhan & Jason Reifler, *When Corrections Fail: The Persistence of Political Misperceptions*, 32 POL. BEHAV. 303 (2010), https://doi.org/10.1007/s11109-010-9112-2 (studying reactions to mock news articles and finding that, in some cases, misleading claims from a politician that were paired with a correction of the misleading claim could sometimes enhance misperceptions among participants, a backfire effect because the correction was designed to reduce misperceptions). However, it is worth noting that subsequent investigation has demonstrated cases where this backfire effect fails to replicate. *See* Thomas Wood & Ethen Porter, *The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence*, 41 POL. BEHAV. 135 (2019), https://link.springer.com/article/10.1007/s11109-018-9443-y (testing 50 scenarios in a series of experiments involving 10,000 participants, the authors found no case of a backfire in which factual corrections increase misperceptions).

102. *See, e.g.*, Bail et al., *supra* note 50 (finding that following a Twitter bot from an opposed political viewpoint could exacerbate political polarization, and that Republicans in particular became significantly more conservative after the treatment).

103 *See id.* at 9217 (finding that Republicans became more conservative after exposure to a liberal-leaning Twitter bot, while Democrats did not demonstrate a statistically significant shift to more liberal positions after similar exposure to a conservative-leaning Twitter bot).

"Our third preregistered hypothesis is that backfire effects will be more likely to occur among conservatives than liberals. This hypothesis builds upon recent studies that indicate conservatives hold values that prioritize certainty and tradition, whereas liberals value change and diversity (40, 41). We also build upon recent studies in cultural sociology that examine the deeper cultural schemas and narratives that create and sustain such value differences (34, 26). Finally, we also build upon studies that observe asymmetric polarization in roll call voting wherein Republicans have become substantially more conservative whereas Democrats exhibit little or no increase in liberal voting positions (42). Although a number of studies have found evidence of this trend, we are not aware of any that examine such dynamics among the broader public—and on social media in particular." *Id.*

are strongly tied to Republican politicians,[104] creating an obvious potential point of partisan difference in responses to the Constitutional Information. The next analysis therefore looks to a possible role of partisan identification in the backfire effects caused by the Constitutional Information. As shown in Figure 3, there is evidence for an interaction effect between partisan affiliation and Constitutional Information consistent with a backfire effect caused by the training.




104. For examples of implicit propagation of misinformation about the First Amendment consider examples cited by Sanders, *supra* note 52, in a discussion of examples of state legislatures passing speech restrictions that are clearly unconstitutional. The examples Sanders cites, from Texas, Arizona, and Florida, all derived from state legislatures currently dominated by the Republican party as determined by checking each state's legislative body composition on ballotpedia.org. *Id.* For examples of explicit propagation of misinformation about the First Amendment, see *supra* note 15 providing instances of Republican politicians incorrectly claiming First Amendment infringements in the case of private entities electing not to facilitate their speech. But for an argument that Democratic politicians are also designing clearly unconstitutional speech-restricting laws, see Mike Masnick, *NY Senator Proposes Ridiculously Unconstitutional Social Media Law That Is the Mirror Opposite of Equally Unconstitutional Laws in Florida & Texas*, TECHDIRT (Jan. 3, 2022, 9:20 AM), https://www.techdirt.com/2022/01/03/ny-senator-proposes-ridiculously-unconstitutional-social-media-law-that-is-mirror-opposite-equally-unconstitutional-laws/.

**Figure 3: Exposure to the Constitutional Information reduces Republicans' support for the content moderation decision. Error bars represent +/- standard error.**

To assess the strength of this potential effect, an interaction term (Republican * Constitutional Information) was included in a least-squares linear regression alongside demographic variables.[105] The results are presented in Table 3. The only statistically significant coefficients are those for age and for the interaction term (Republican * Constitutional Information). Gender, race, political party, and exposure to the Constitutional Information do not significantly predict support for the platform's decision to suspend the user.

| **Variable** | **Coefficient (SE)** |
|---|---|
| Constant | **4.3***** |
| | (.19) |
| Gender (Female) | .11 |
| | (.10) |
| Race (White) | -.16 |
| | (.12) |
| Age | **.009**** |
| | (.003) |
| Republican | -.24 |
| | (.18) |
| Constitutional Information | -.06 |
| | (.12) |
| Republican * Constitutional Information | **-.64**** |
| | (.23) |
| * p < .01, ** p < .001, *** p < .0001 | |

**Table 3: Least-squares linear regression coefficients when predicting support for platform decision.**

105. The linear regression to assess the influence of demographic and political factors was pre-registered, but the inclusion of the interaction term was post hoc.

The significance of age in predicting support for content moderation is consistent with the results of a 2021 study by Riedl, which found a significant positive coefficient for age in a similarly large and representative sample of U.S. adults.[106] Likewise, the lack of significance for gender and race are also consistent with Riedl's results.[107]

The interaction term Republican * Constitutional Information has a large, negative coefficient (-0.64). An age effect of equal magnitude would require seven decades (70 * .009 = .63). The Republican-training interaction effect—which creates the backfire effect—is therefore by far the most impactful explanatory variable identified by the linear regression. The importance of partisan identity in the backfire effect is consistent with previously cited work identifying backfire effects specifically in Republicans.[108]

This isn't merely an interesting experimental finding, but one that has real potential consequences. For those who seek to correct misrepresentations about First Amendment protections - such as those seemingly strategically deployed by opportunistic politicians - such actors should understand that anti-misinformation efforts may have unintended consequences beyond directly adjusting the degree of First Amendment literacy in targeted populations. That is, proponents of First Amendment literacy efforts should know that their efforts could have unintended consequences, such as in the case of support for content moderation. Likewise, those who seek to instrumentalize empirical connections between formally unrelated concepts, such as the connection identified in this Article between knowledge of the First Amendment and support for private content moderation policies, should be aware that such efforts can backfire.

In summary, in addition to documenting a backfire of the educational intervention, the results demonstrate a partisan source for this backfire. As shown in the linear regression, Republicans who received the Constitutional Information exhibited a mean decrease in support of -0.64 points (on a Likert scale of 5, thus a substantial drop). This experimental finding provides evidence of a highly fraught situation. Misunderstandings

106. Riedl et al., *supra* note 21, at 9–10.

107. *Id.* at 10.

108. *See generally*, Bail et al., *supra* note 50 and their discussion regarding their hypotheses.

of law can create challenging contours of public opinion, which digital platforms and lawmakers must navigate carefully, regardless of their ultimate policy objectives.

### *E. Constitutional Information with a Governmental Emphasis*

One fair question is whether something about the Constitutional Information is unusual such that the backfire result is unlikely to generalize. Might there be something specific about the wording of the Constitutional Information treatment that creates a backfire effect in Republicans?[109]

To understand the extent to which the specific wording of the training might create the effect, it is worth exploring alternative forms of constitutional information. In the experiment, a fourth informational treatment was run simultaneously in Stage 1[110] (the Government Constitutional Information) to test an alternative version of constitutional information.[111] This alternative version of the information conveyed similar information to the original treatment, but with an emphasis on the limitations of the First Amendment restrictions to government (rather than to the lack of limitations on private entities).

The results with respect to the Government Constitutional Information differed in some interesting ways. The correctness rate on a constitutional question with government emphasis went down ($p < .05$) with training (36 percent correctness rate in the Government Knowledge Elicitation[112] condition as compared to 26 percent correctness rate in the Government

109. One could likely formulate a variety of possible theories to explain the backfire effect. The elucidation of a causal mechanism is left to future work.

110. This was not included in the initial presentation of the design for the sake of simplicity. However, the treatment was implemented in the original experiment and described in the pre-registration materials.

111. The full text of this treatment is provided in the online supplementary materials. Nielsen, *Supplements*, *supra* note 70. While the original training emphasized the lack of applicability of First Amendment restrictions to private entities, the Government Constitutional Information emphasized the applicability of First Amendment restrictions to governmental entities.

112. This mirrored the original Knowledge Elicitation condition, but the question for participants was about a public university, rather than a private university, preventing students from inviting a controversial speaker to campus.

Constitutional Information).[113] Also, those who answered the constitutional question correctly were less rather than more supportive of content moderation ($p < .0001$).[114] Thus, in this version of the training neither H1 ( Constitutional Content Moderation Connection) nor H2 (the Correctable Belief hypothesis) held in the predicted direction. Both effects were significant but ran contrary to what was predicted.[115]

Yet, despite these differences from the pattern for the original training, the backfire effect still occurred, in contravention of H3 (the Connection Manipulation hypothesis). Participants who received the Government Constitutional Information were less supportive of content moderation than those in the Control Group condition, with mean support of 4.2 and 4.5, respectively ($p < .01$). And, as with the Constitutional Information, there was an interaction effect between identifying as a Republican and exposure to Government Constitutional Information (post hoc, $p < .001$). Thus, the backfire effect was not limited to one possible form of Constitutional Information. Also, the backfire effect continued to manifest specifically in Republicans, even with a different emphasis in the information provided to participants. This finding suggests that a backfire effect is not an artifact of the original Constitutional Information design, but rather is likely to generalize.

---

113. Interestingly, this suggests the possibility of a backfire effect even with respect to increasing constitutional knowledge. Or, an alternative interpretation, discussed earlier, is that responses to the constitutional questions can be taken as expressions of normative beliefs or, alternately, as an expression of desire for legal evolution rather than as an expression of belief about the state of the law. However, as described earlier, these interpretations seem unlikely to obtain given results showing differences in responses to normative as compared to descriptive variables collected in Stage 3 in the case of Constitutional Information. These results showed that the Constitutional Information affected only the participants' descriptive beliefs, not their normative beliefs.

114. *See* the online supplementary materials at Nielsen, *Supplements*, *supra* note 70.

115. Those choosing the correct answer to the constitutional question would have indicated that yes, there could be a constitutional infringement (because the question asked about a public rather than a private university), potentially emphasizing the rights of individuals' speech. Thus, it is not very surprising that those answering the constitutional question correctly in the case of a public university, rather than a private university, might tend to evince more support for individual rights and so possibly less support for content moderation. Thus the failure of H1 to hold is not very surprising. What is surprising is the failure of H2 to hold - that is, it's surprising that the Government Constitutional Information may even have had a backfire effect on accurately answering the constitutional question.

### *F. Support for a User's Constitutional Lawsuit*

Another fair question is whether support for content moderation reduces to a single dimension, as implied in the analysis so far presented. In the preceding analyses, the basis for assessing attitudes towards content moderation was expressed support for a company's decision to suspend a user's account.

But the analysis presented so far does not provide full information regarding participants' opinions about content moderation generally, and could be consistent with structures of support distinct from a simple, unidimensional model. For example, it could be that participants support the account suspension in the vignette, but also support the user's constitutional lawsuit to fight that account suspension. That is, among many possibilities, perhaps supporting account suspension is simply an expression of support for vigorous action on behalf of any person or entity supporting their beliefs (such an attitude would be consistent with one generally favorable to a clamorous marketplace of ideas), rather than support trending exclusively in the favor of supporting companies' actions to regulate speech on their platforms. By gauging the level of support for the banned user's constitutional lawsuit, it is possible to examine a related, but distinct, judgment about content moderation as compared to the one examined so far, which emphasized the platform's prerogatives.

The relationship between supporting the platform's content moderation decision and supporting the user's constitutional lawsuit was strong, with a correlation of -.6. In other words, those who indicated high support for the content moderation decision tended to indicate low support for the constitutional lawsuit, and vice versa. The measure of support for the user's constitutional lawsuit also correlated with performance on the constitutional law question, with a mean support for the lawsuit of 1.3 for those who answered correctly, and 2.1 for those who answered incorrectly, in the Knowledge Elicitation condition (post hoc, $p < .0001$), consistent with H1, the Constitutional Content Moderation Connection hypothesis.

There was, however, no backfire effect of the training on support for the user's constitutional lawsuit (in contrast to the backfire effect in support for the platform's decision). Those in the Constitutional Information treatment evinced the same low mean support level for the constitutional lawsuit as those in the Control Group condition (1.8). Likewise, in a least-squares linear regression, there was no interaction effect between Republican identity and exposure to the training. There was, however, an effect of Republican identity ($p < .05$) and also of age ($p < .001$).

In summary, correct responses to the constitutional law question were correlated with lower support for the user's constitutional lawsuit (H1), consistent with the earlier reported results. However, there was no backfire effect of the training on support for the lawsuit (H3). These results suggest that different parties to a content moderation dispute will likely face somewhat different sets of considerations when inferring how their actions will likely be judged by ordinary Americans.

### *G. Limitations*

This experiment – of course – cannot address all questions related to folk beliefs about the scope of First Amendment and ramifications for lay judgments of content moderation. A few key limitations are addressed here.

A prime limitation is whether participants should be described as having "inaccurate" beliefs about the scope of constitutional protections. Perhaps participants who answered the constitutional law question incorrectly were expressing a political stance, not failing to correctly answer a constitutional law question.

But even if true, this alternate interpretation does not undermine the analysis, it merely changes the label. That is, one could possibly label such participants as those who disagree with the law or who believe that private universities ought to be bound by the First Amendment, rather than describing them as holding inaccurate beliefs. Also, while this alternative understanding is not problematic to the results, there is evidence from the experiment that such an interpretation is probably not more likely to be a true description of what is happening in participants' minds when they respond to the questions in the experiment. Consider the robustness checks in Stage 3, which showed that participants changed their responses to descriptive but not normative questions about the scope of constitutional speech protections. It seems unlikely that participants would choose to answer a descriptive question as normative one in Stage 1, but not then do so in Stage 3. Thus the robustness check suggests that participants did respond to the Stage 1 question as a descriptive question.

Another potential limitation is that the experiment focuses on one specific content moderation scenario. It is possible that the responses to this

scenario might not replicate over other scenarios of potential interest.[116] But the optional freeform responses that participants submitted tended to address the topic of content moderation generally rather than to focus on the specific fact of a user account suspension, providing suggestive evidence that many different ways of implementing content moderation policies are likely to be correlated in the minds of laypeople. Future empirical work, however, could look at nuances in how laypeople respond to the wide variety of content moderation actions available to digital platforms.[117]

Another limitation is that this experiment looks at attitudes given the existing political context: one in which the government is not involved in guiding platforms' online speech regulation. It is possible that a different level of government involvement or transparency by platforms could result in quite different judgments or a different empirical connection between beliefs about the First Amendment and support for platforms' content moderation decisions. For example, if the government were to take a more active role in defining permissible content moderation practices or in assuring review of decisions,[118] it is possible that the relationship between constitutional beliefs and attitudes about content moderation might be

116. Including, but not limited to, comment deletion, post deletion, restrictions on sharing, mandatory labeling, and responses to other forms of media such as video or audio.

117. There are many different ways in which content moderation can be varied. In the case here, content moderation occurs at the level of the user, permanently banning a particular user as a way to stop problematic content. It is also, of course, possible to moderate content directly without taking action against a particular user account. For example, even with respect to timing, there are many ways in which platforms can react to problematic content, including ex ante content moderation, ex post content moderation, and ex post reactive content moderation. For a description of some of the variety in automated content moderation (which is not the only form of content moderation), see Spandana Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, NEWAMERICA (July 22, 2019), https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/ (detailing these three options regarding timing of content moderation in the Introduction).

118. Assuming, arguendo, that this is even permissible under the First Amendment.

weakened, say if participants perceived some other statute as a stronger and more relevant source of legal protection.

Regarding the backfire effect, it is not clear the extent to which the effect may be the particular result of the mechanism of delivery or perceived identity of the source of the Constitutional Information. Participants knew they were participating in an academic experiment. Republican politicians and even ordinary Republicans seem, increasingly, to see themselves as targeted by or otherwise opposed by academic researchers, who are perceived to be overwhelmingly liberal.[119] It is therefore possible that Republican research subjects particularly distrusted the Constitutional Information in this experiment, thus providing one potential causal mechanism for the backfire effect. However, this possibility does not challenge the external validity of the finding as it seems, at least for the immediate future, that potential providers of such messaging in the real world would likely also be perceived as liberal.

Finally, there are two ways in which the sampling is limited. First, the proportion of self-identified Republicans in the experiment sample was lower than it would be in a truly representative sample of Americans. But there is no reason to think that the direction of the effects would be different. If anything, a more representative sample would include more Republicans and would therefore likely show larger effects of the backfire at the population level.

A second concern about the sampling is that the participants—who have sought out an opportunity to answer survey questions online—are probably more technologically proficient and more interested in technology issues, such as online speech regulation, than is the general public. This might explain the surprisingly high level of support for the firm's decision in the vignette study, or the relatively high levels of sophistication evinced in some optional freeform responses. But this sampling skew would not create the specific effects of interest here, specifically the constitutional content moderation connection and the backfire effect.

The limitations identified here point to the need for more work to understand the probability of a backfire effect in the real world, and the causal mechanisms for such an effect. However, such limitations do not undercut the contributions of this Article, namely (1) that there is an

119. *See* Graham Vyse, *Liberals Can't Ignore the Right's Hatred for Academia*, NEW REPUBLIC (July 13, 2017), https://newrepublic.com/article/143844/liberals-cant-ignore-rights-hatred-academia.

empirical connection in lay beliefs between distinct areas of law and policy and (2) that literacy efforts directed at the First Amendment could potentially have surprising consequences as a result of this connection.

## CONCLUSION

On its face, online speech regulation may, at least for now,[120] be a matter of private law. Nonetheless, laypeople draw upon far broader legal notions in their own assessments of content moderation. This Article shows that the ongoing and controversial public debate about appropriate online speech governance may have deep perceived constitutional roots that need to be addressed carefully, particularly given the possibility for backfire effects.

This Article identifies two key findings that are surprising and merit recapitulation. First, a majority of ordinary Americans in the representative sample of U.S. adults indicated incorrect beliefs regarding the breadth of First Amendment protections, with such beliefs correlated to lower support for a platform's decision to enforce its online speech policies. Second, constitutional information designed to address these legal misperceptions backfired, specifically among Republicans.

The experiment has established that constitutional misapprehensions are common and likely have real world consequences for technology policy. The experimental findings may come as no surprise to technology firms. Such entities have sophisticated data analytics and full-time researchers whose jobs are dependent on understanding how people think and how they will respond to various, nuanced manipulations of products or policies.

But extant scholarly work has done very little in the way of encouraging or undertaking empirical investigation in this area. This work contributes to a small but growing body of quantitative experimental research on fundamental rights.[121] This work thus ultimately serves as an

---

120. *See, e.g.*, Press Release, Richard Blumenthal, U.S. Sen. for Conn., U.S. S., Blumenthal & Blackburn Introduce Comprehensive Kids' Online Safety Legislation (Feb. 16, 2022), https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-and-blackburn-introduce-comprehensive-kids-online-safety-legislation (2022 will likely be a year of active regulatory proposals if not of new legislation).

121. *See generally* Adi Leibovitch & Alexander Stremitzer, *Experimental Methods in Constitutional Law*, U. CHI. L. REV. ONLINE (Apr. 5, 2021),

example and an argument in favor of more such empirical work, and the relevance of such work to technology policy. Legislators and regulators need to do more to understand the views of ordinary people, views which can be complex and which can deviate from legal realities in important ways.

It is fair to characterize the main result of this experiment as one of possibility: It is possible that attempting to increase support for content moderation through legal education could backfire. It is possible that efforts to counter legal misinformation could have unanticipated effects beyond the direct educational targets of such campaigns. Finally, it is possible that *beliefs about* the First Amendment constitute as much of a significant factor (or even roadblock) as the First Amendment itself[122] when crafting content moderation policy to stand up to the rigors of American expectations and desires.

Lawmakers are at a potential crossroads regarding whether and how to change the governance model for online speech. No doubt, content moderation policy is a complicated topic, worthy of the attention bestowed upon it from legal theorists and sophisticated technology companies. However, these elite entities have dominated the discussion for too long. It's time to hear more from ordinary people, and learn how they judge the rights and wrongs of online speech freedom.

---

https://lawreviewblog.uchicago.edu/2021/04/05/cv-leibovitch-stremitzer/ (discussing the importance of quantitative experimental research in constitutional law scholarship).

122. *See generally* RICHARD L. HASEN, CHEAP SPEECH: HOW DISINFORMATION POISONS OUR POLITICS – AND HOW TO CURE IT (2022) (evincing the view that the First Amendment substantially limits the range of permissible solutions available in the United States for tackling the challenges of easy and cheap speech opportunities on social media).