Harvard Journal of Law & Technology
Volume 35, Number 1 Fall 2021

THE MYTH OF THE CHILLING EFFECT

*Suneal Bedi**

ABSTRACT

The chilling effect — historically associated with protecting First Amendment rights — has more recently become a tool used to argue against social media platform (e.g., Facebook, Twitter) policies of restricting content. The 70-year-old principle invalidates regulations if censoring the unwanted speech will also deter or "chill" related but valuable speech. Despite its longstanding use and recent significance, little to no empirical work has been done on whether this phenomenon exists. This Article finally fills this gap. It employs an empirical study using social media speech restrictions analyzed with text analysis to conclude that, in the social media context, the chilling effect has little to no impact on the content of the message; at most, it subtly alters the specific style or tone used. This study therefore confirms the existence of the phenomenon but simultaneously raises serious concerns about its usefulness as a guiding constitutional legal principle when assessing speech regulations.

* Assistant Professor of Business Law, Indiana University. Affiliated Faculty, The Center for Law, Culture, and Society at Indiana University Maurer School of Law. I would like to thank the following parties for helpful conversations, suggestions, and comments: Mark Tushnet, Leslie Kendrick, Alexandra Roberts, Matt Turk, Sonu Bedi, Monu Bedi, Vikram Bhargava, Dave Reibstein, Ike Silver, Rom Schrift, Todd Haugh, and the discussants at the 2021 Law & Society Conference. Thanks to Lin Ye and Wei-Chung Lin for great research and editing assistance. Finally, thank you to the journal editors at the Harvard Journal of Law & Technology for great comments and suggestions.

TABLE OF CONTENTS

I. INTRODUCTION

A guiding legal principle of the First Amendment is the existence of the chilling effect. That is, certain speech regulations may have the "indirect effect of deterring a speaker from exercising her First Amendment rights."[1] Put another way, a regulation may only seek to censor, limit, or restrict certain types of speech, but because citizens fear sanctions, the same regulation ends up censoring, limiting, or restricting other potentially valuable types of speech.[2]

---

1. Monica Youn, *The Chilling Effect and the Problem of Private Action*, 66 VAND. L. REV. 1473, 1474 (2013).

2. For discussions on what the chilling effect is and how it has been utilized in legal jurisprudence, see generally Frederick Schauer, *Fear, Risk and the First Amendment: Unraveling the "Chilling Effect,"* 58 B.U. L. REV. 685 (1978) (discussing what the chilling effect is and how it has been used in court cases). *See also* Leslie Kendrick, *Speech, Intent, and the Chilling Effect*, 54 WM. & MARY L. REV. 1633, 1640–41 (2013) (discussing the aspect of speaker intent as an important but often overlooked part of the chilling effect); Henry P. Monaghan, *Constitutional Fact Review*, 85 COLUM. L. REV. 229, 268 (1985) ("The familiar 'chilling effect' rhetoric asserts that first amendment values are very fragile and especially vulnerable to an 'intolerable' level of deterrence; and the danger of impermissible deterrence is real . . . ."); Daniel J. Solove, *The First Amendment as Criminal Procedure*, 82 N.Y.U. L. REV. 112, 142–43 (2007) (discussing the deterrent effects of speech regulation in the context of criminal law); Margot E. Kaminski & Shane Witnov, *The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech*, 49 U. RICH. L. REV. 465, 483–

The specter of the principle is often raised in the face of broad and potentially ambiguous restrictions on speech that are argued to be unconstitutional, closely related to constitutional issues of the overbreadth doctrine.[3] The theory is if citizens are not confident of exactly what speech is being limited, they may overregulate their own speech for fear of sanctions, and hence may chill (deter and change) their speech in an unnecessary way.[4]

Although historically associated with constitutional issues surrounding the First Amendment and government censorship, the chilling effect principle has been recently employed against private speech regulations. The debate around limiting hate speech, offensive language, and political conspiracies on social media sites like Facebook and Twitter has reinvigorated discussions of the chilling effect. In January 2021, both Facebook and Twitter blocked users including former President Donald Trump and conspiracy theory supporters from their platforms, and publicly removed posts which were perceived to contribute to propaganda and violence.[5] This led to both criticism and praise, with some arguing that these social media speech restrictions problematically deter and chill otherwise valuable speech, and others arguing that no chilling takes place, and even if it did, it should not matter because the restrictions are important and necessary.[6]

The chilling effect is so frequently used in court cases, scholarly articles, and news outlets that its existence has become self-evident to

---

93 (2015) (discussing investigations into how surveillance contributes to the chilling effect); Mark Tushnet, *Why Protect Falsity?*, JOTWELL (Dec. 20, 2010), https://conlaw.jotwell.com/why-protect-falsity/ [https://perma.cc/AM3G-HU4G] (arguing that the chilling effect has been absent in discussions of false statements of facts).

3. *See* Schauer, *supra* note 2, at 685–86, 685 n.7.

4. *See id.* at 695. *See generally* Kendrick, *supra* note 2.

5. For a detailed press release of the ban, see *Permanent Suspension of* @real*DonaldTrump*, TWITTER BLOG (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension [https://perma.cc/DE6A-WUCQ]. *See also* Todd Spangler, *Facebook Bans Massive Network of Fake Accounts That Were Spreading Pro-Trump Propaganda*, VARIETY (Dec. 20, 2019, 2:14 PM), https://variety.com/2019/digital/news/facebook-bans-the-bl-pro-trump-propaganda-1203450114/ [https://perma.cc/PZ7H-2K99].

6. For discussion on how social media speech restrictions affect speech, see Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011 (2018) (noting that social media sites are engaging in speech censorship in the face of external pressures). *See also* Marjorie Heins, *The Brave New World of Social Media Censorship*, 127 HARV. L. REV. F. 325, 325–26 (2013) (detailing the extent to which social media companies end up limiting otherwise protected First Amendment speech); Kerry Flynn, *After Charlottesville, Tech Companies Are Forced to Take Action Against Hate Speech*, MASHABLE (Aug. 16, 2017), https://mashable.com/2017/08/16/after-charlottesville-tech-companies-action-nazis/ [https://perma.cc/AS9Z-Q6ZM] ("Facebook, Google, Spotify, Uber, Squarespace, and a variety of other tech companies are taking action to curb the use of their platforms and services by far-right organizations."); John Koetsier, *Social Censorship: Should Social Media's Policy Be Free Speech?*, FORBES (Oct. 25, 2020, 10:47 PM), https://www.forbes.com/sites/johnkoetsier/2020/10/25/social-censorship-should-social-medias-policy-be-free-speech/ [https://perma.cc/JB85-EZ3M] (discussing arguments against social media censorship).

the legal community — no law school First Amendment course is complete without some detailed discussion of the chilling effect.[7] Yet, surprisingly, little or no empirical scholarship has grappled with whether the chilling effect exists. Does speech in the face of censorship actually get chilled? If it does, in what ways is it chilled?

These questions sit at the core of this common constitutional guiding principle. However, they have effectively gone unanswered for over seventy years: this legal principle is "founded . . . on nothing more than unpersuasive empirical guesswork."[8] Measuring whether and how speech gets chilled will not only help legal scholars and judges better understand which regulations to question but also will foster more informed decision-making when private companies attempt to censor speech.

Measuring the chilling effect is difficult for various reasons including data collection, data analysis, causal events, and confounding factors.[9] In addition, various types of speech activities (e.g., more factual reporting-based, opinion-based, or a mix of the two) and mediums of speech activity (e.g., social media, in-person protests, traditional news reporting) may manifest the effect differently. A leading scholar once commented that the human behaviors that underly the chilling effect are "most likely unprovable."[10] Therefore, it is not terribly surprising that extant scholarship has neither confirmed nor disproved its existence.[11]

This Article seeks to fill this gap. It is the first empirical study analyzing speech content and measuring whether and how speech is actually chilled. It focuses on social media speech activity — online restaurant reviews — that has both fact and opinion components. In this social media context, this Article empirically showed that the chilling effect created little to no impact on the content of the message; at most, it slightly altered the specific style or tone used.

---

7. Most First Amendment casebooks discuss "chilling speech" as a fundamental theoretical perspective. *See, e.g.*, RUTHANN ROBSON, FIRST AMENDMENT: CASES, CONTROVERSIES, AND CONTEXTS 6 (2d ed. 2020) ("Concepts such as 'chilling speech' or 'secondary effects' waver between theory and doctrine."). *See also* Schauer, *supra* note 2, at 685 ("[T]he chilling effect doctrine underlies the resolution of many cases in which it is neither expressed nor clearly implied.").

8. Kendrick, *supra* note 2, at 1684. For a discussion of the lack of empirical rigor associated with the chilling effect, see Kendrick, *supra* note 2, at 1681–84.

9. *See* Kendrick, *supra* note 2, at 1675–80.

10. *See* Schauer, *supra* note 2, at 730.

11. The existing empirical work on the subject has fallen short, as described further below in Part III. *See, e.g.*, Jonathon W. Penney, *Internet Surveillance, Regulation, and Chilling Effects Online: A Comparative Case Study*, 6 INTERNET POL'Y REV., May 26, 2017, at 1. One article attempted to measure how behavior extending beyond speech can get chilled. Brandice Canes-Wrone & Michael C. Dorf, *Measuring the Chilling Effect*, 90 N.Y.U. L. REV. 1095, 1097–98 (2015) (using an event study to show that changes in policy affecting abortion timelines did have a chilling effect on behavior).

To do this, the Article uses an actual (not hypothetical) private social media speech restriction and experimentally manipulates how broad or narrow the speech restriction is. It then analyzes the online speech output itself using text analysis and third-party evaluations to show whether and how exactly the speech is changed in the face of a given restriction. The results of this empirical study show that social media user speech is more affected as regulations become broader — which confirms the existence of the chilling effect. But, while users did change their speech, the overall message of the speech did not change at all. That is, users change the tone of what they say (in this case by making it more positive) in the face of restrictions, but ultimately communicate the same message.

This result both partially confirms the chilling effect and simultaneously raises serious issues about how it is has been historically treated as a legal principle. While people do change how they speak in the face of a regulation, restrictions do not seem to have a substantive effect on the message communicated at least in the context studied. The lack of an observable and substantive chilling effect calls into question whether the chilling effect is a principle courts, litigants, governments, and private companies should put so much emphasis on.

In addition to questioning whether the chilling effect should be a guiding principle, the results of this Article may point to the conclusion that the chilling effect can be a good thing in some contexts. By changing only the tone of how individuals speak, rather than the actual content or purpose of a message, the chilling effect can encourage more civil forms of speech that are less offensive. When individuals are less offended, they are more likely to engage in thoughtful exchanges. These exchanges then ultimately promote participation in speech activity rather than upset it. This Article then points to the insight that recent restrictions on social media platforms may ultimately be good policy not in spite of, but because of the chilling effect.

This Article is in five Parts. Part II describes the current state of the legal principle of the chilling effect focusing on its recent use in private social media speech restrictions and extant empirical work. Part III introduces a novel empirical study that measures whether and how speech is chilled in the face of a private social media speech regulation. Part IV discusses the implications of the results for both public and private regulations going forward, focusing on the potential for chilling to actually promote speech rather than frustrate it. Part V concludes.

II. THE CURRENT STATE OF THE CHILLING EFFECT

A. Historical Background

The original concept of "chilling" dates back to 1952 with the Supreme Court case *Wieman v. Updegraff*.[12] In that case, Oklahoma required each state employee to pledge a "loyalty oath" indicating that they currently were not and had not been for five years part of any organization that the Attorney General of the United States deemed to be "subversive."[13] The Court invalidated the statute as it violated the Fourteenth Amendment. Justice Frankfurter in his concurring opinion stated: "Such unwarranted inhibition . . . has an unmistakable tendency to chill that free play of the spirit which all teachers ought especially to cultivate and practice; it makes for caution and timidity in their associations by potential teachers."[14] Justice Frankfurter's theory is that teachers would not associate freely with various organizations that may be protected and valuable associations because they may fear government sanctions.[15]

The "chilling effect" later found its way into the speech regulation jurisprudence in 1963 in *Gibson v. Florida*.[16] Subsequently, the concept of chilling has been extensively applied to invalidate various government regulations, in particular those that implicate the First Amendment.[17] The idea is that the phenomenon of chilling can serve as a weapon to argue against the potential negative externalities associated with various government regulations.[18]

In terms of a simple, workable definition, leading First Amendment scholar Frederick Schauer defines the chilling effect as something that occurs "when individuals seeking to engage in activity protected by the [F]irst [A]mendment are deterred from so doing by governmental regulation not specifically directed at that protected activity."[19] Take for example a federal law that makes it illegal to say anything vulgar to

12. Wieman v. Updegraff, 344 U.S. 183, 195 (1952).

13. *Id.* at 183.

14. *Id.* at 195 (Frankfurter, J., concurring).

15. *Id.* at 195–96 (Frankfurter, J., concurring).

16. Gibson v. Fla. Legis. Investigation Comm., 372 U.S. 539, 556–57 (1963) ("While . . . all legitimate organizations are the beneficiaries of these protections, they are all the more essential here, where the challenged privacy is that of persons espousing beliefs already unpopular with their neighbors and the deterrent and 'chilling' effect on the free exercise of constitutionally enshrined rights of free speech, expression, and association is consequently the more immediate and substantial.").

17. In 1967, Justice Harlan claimed that the chilling effect doctrine had become "ubiquitous." Zwickler v. Koota, 389 U.S. 241, 256 n.2 (1967) (Harlan, J., concurring).

18. *See generally* Schauer, *supra* note 2 (providing a detailed background on the state of the chilling effect as it related to speech in 1978).

19. *Id.* at 693.

politicians.[20] In theory, this law would chill speech because individuals may be deterred from saying critical things to politicians, that would otherwise be protected by the First Amendment, due to the fear of being labeled vulgar and hence may change or taper their speech in certain ways.

There are two important parts to the traditional definition of the chilling effect that are worth discussing further: deterrence and governmental regulation.

The chilling effect principle rests on the assumption that individuals are frequently deterred in the face of government regulations.[21] But why are people deterred from engaging in an activity by a law that makes it illegal to engage in a different activity? Scholars have recognized that individuals are fearful of criminal and civil sanctions.[22] Ideally, of course, individuals would know exactly what would create a sanction and would be able to easily adapt their speech to the constraints of a regulation.[23]

However, this is often not the case with speech regulations. The deterrence aspect of the chilling effect is particularly highlighted when a regulation's application is uncertain. The uncertainty of whether one's speech is covered under a regulation[24] makes a speaker more likely to overcorrect with the intent to confidently speak in a legal manner. This behavior is commonly found in many decision-making contexts.[25] Therefore, we might predict that broader and more vague

---

20. This is reminiscent of the premise of *N.Y. Times Co. v. Sullivan*, 376 U.S. 254 (1964), in which the Court restricted the ability of public officials to sue for defamation.

21. Schauer, *supra* note 2, at 689 ("The very essence of a chilling effect is an act of deterrence."). *See also* Freedman v. Maryland, 380 U.S. 51, 59 (1965); *Gibson*, 372 U.S. at 557.

22. Schauer, *supra* note 2, at 694 ("The answer must be that these individuals fear punishment or other detriment in spite of the lawful nature of their contemplated behavior.").

23. Schauer explains it well: "In an ideal world, there would be neither error nor uncertainty in the legal process. . . . As a result of the foregoing . . . . [a]ny individual could instantly and effortlessly ascertain whether his contemplated conduct would be a violation of a given enactment . . . ." Schauer, *supra* note 2, at 694.

24. *See* Kendrick, *supra* note 2, at 1652–55 ("Uncertainty may stem from ambiguous rules or erroneous applications. Either of these may make a speaker fear that he will be held liable for speech that should properly be protected. The closer his speech is to the line between protected and unprotected, the more pronounced this uncertainty will be.").

25. The decision-making literature is vast, and in this context, focuses on how risk-averse individuals make decisions that seem to overcorrect for a given constraint. *See, e.g.*, Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decision Under Risk*, *in* HANDBOOK OF the FUNDAMENTALS OF FINANCIAL DECISION MAKING: PART I 99 (Leonard C. MacLean & William T. Ziemba eds., 2013) (describing prospect theory as a way of understanding decision-making under uncertainty); Richard Craswell & John E. Calfee, *Deterrence and Uncertain Legal Standards*, 2 J.L. ECON. & ORG. 279, 279–80 (1986) (discussing an economic theory of how citizens make decisions under uncertain laws). For deterrence in the face of financial decisions, see Henrik Lando, *Does Wrongful Conviction Lower Deterrence?*, 35 J. LEGAL STUD. 327, 327–37 (2006) (showing that the deterrence effects of wrongful conviction are more complicated than previously thought). For a background on deterrence rooted in cognitive theory, see Jeffrey D. Berejikian, *A Cognitive Theory of Deterrence*, J. PEACE RSCH. 165 (2002).

speech regulations, which naturally create more uncertainty, will have a larger chilling effect than speech regulations that are more narrowly tailored.

Although seemingly obvious, deterrence is itself a complicated and messy topic. While on a first read, deterrence simply means not doing something, this may have various meanings in the speech context. It could mean, for example, not engaging in any kind of related speech whatsoever.[26] This is, however, a limited view of what deterrence is. Viewing the concept of deterrence more broadly would lend one to conclude that an individual may still engage in speech, but due to fear of repercussion may change how exactly they speak, what words they use, and what they communicate. This is still in the context of overcorrection given the existence of the regulation.[27] Scholars have noted this point as well: "[t]he chilling effect encompasses not only outright suppression of speech but also subtle modification."[28]

Take for example a law that makes it illegal to say harmful things to your colleagues because it could create depression and psychological harm. Imagine that Anjali, incredibly upset at her colleague Ben, wants to say to Ben that he is a misogynist. Assume the law does not seek to limit this kind of speech. Yet, Anjali may be deterred from making the statement because it could be misconstrued to be harmful, and hence, she would face sanctions. She could be deterred in the traditional sense[29] and not say anything at all to Ben. Or Anjali could still say something to Ben but end up saying something less poignant due to fear of sanctions. Maybe Anjali simply says "Ben, you're not a good guy" — something not clearly harmful but also not what Anjali really wanted to say.

Therefore, we might predict that when faced with speech regulation, individuals could simply decide to not engage in otherwise protected speech or overcorrect and alter their speech in some significant way — in this context, deterrence is best understood as a sliding scale

26. This is the more standard way to think about deterrence and tracks early cases focusing on the chilling effect. *See generally, e.g.*, N.Y. Times Co. v. United States, 403 U.S. 713 (1971); N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964).

27. For example, one interpretation of the *Sullivan* case is that reporters will likely be hesitant to say critical things about politicians and may end up saying something that is nicer than what they would have. In effect, they were deterred from saying what they wanted to in the face of sanctions, so instead said something similar but less poignant.

28. Kendrick, *supra* note 2, at 1678; *accord* Russell L. Weaver & Geoffrey Bennett, *Is the New York Times "Actual Malice" Standard Really Necessary? A Comparative Perspective*, 53 LA. L. REV. 1153, 1189 (1993) (arguing that modification of speech is also chilling).

29. By traditional sense, I mean chilling that manifests in either less speech (e.g., the speaker decides not to speak at all) or less impactful speech (e.g., the communicative effect of the speech is changed).

(from zero speech on one end to some, albeit demonstrably changed, speech on the other).[30]

The original conception of the chilling effect stated above was developed in the context of public regulations,[31] which are government (either federal, state, or local) regulations that sought to limit how individuals could express themselves. After all, the First Amendment by its very nature implicates only government action.[32]

Potentially one of the foremost landmark First Amendment cases was *New York Times v. Sullivan*.[33] In *Sullivan*, the New York Times published an advertisement of supporters of Martin Luther King criticizing how the Alabama police treated civil rights protestors.[34] The Alabama court system found that there were some inaccuracies in the advertisement and hence it was per se defamatory to the Alabama police.[35] The Supreme Court, however, overturned the verdict, because it found that the Alabama defamation law was too restrictive and thus amounted to a violation of the First Amendment.[36]

In doing this, the Court held that false statements are inevitable in free debate and that they must be protected if the freedoms of expression are to have the "breathing space" that they "need . . . to survive."[37] Critical to the Court's decision was the fact that a per se defamation law would deter individuals from engaging in otherwise protected speech. In his concurrence, Justice Goldberg stated that the "opinion of the Court conclusively demonstrates the chilling effect of the Alabama libel laws on First Amendment freedoms in the area of race relations."[38] If public officials can forestall criticism of their official conduct by resorting to friendly juries, criticism by the press and citizen will be silenced.

Further extensions of the chilling effect can be seen in cases involving simple lies. In *United States v. Alvarez*, the Court found that the Stolen Valor Act violated the First Amendment.[39] The Stolen Valor Act made it unlawful for individuals to lie about receiving military

30. The point I make here is an important one. I do not claim that speech is chilled simply when it is changed; in some ways, the point of regulation is to change speech. However, speech can be chilled if it is changed in a way that deters an individual from saying something that is protected because of fear that it may be perceived as illegal.

31. Schauer's whole focus in his seminal article on the chilling effect is on public regulations that seek to limit speech activity. *See generally* Schauer, *supra* note 2.

32. The First Amendment does not explicitly contemplate private actions but rather addresses congressional actions. Of course, the Fourteenth Amendment then made it applicable to state congressional actions as well.

33. N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964).

34. *Id.* at 257–58.

35. *Id.* at 262–64.

36. *Id.* at 264.

37. *Id.* at 272 (quoting NAACP v. Button, 371 U.S. 415, 433 (1963)).

38. *Id.* at 300–01 (Goldberg, J., concurring).

39. United States v. Alvarez, 567 U.S. 709, 730 (2012).

accolades.[40] The Court, in part relying upon the chilling effect, found that the law as written "would give government a broad censorial power" and that the "mere potential for the exercise of that power casts a chill."[41] Also, "the threat of criminal prosecution for making a false statement can inhibit the speaker from making true statements, thereby 'chilling'" valuable speech.[42] Much like in *Sullivan* and previous cases, the Court, in finding the statute was unconstitutional, held that imposing a mens rea requirement for false statements would create "'breathing room' for more valuable speech by reducing an honest speaker's fear that he may accidentally incur liability for speaking."[43]

The chilling effect has been applied widely to argue against public government regulations in various categories such as obscenity,[44] pornography,[45] fraud,[46] general commercial speech,[47] privacy,[48] the intentional infliction of emotional distress,[49] and even campaign funding.[50] In addition, the possibility of government regulations chilling speech has been the subject of many law reviews.[51]

---

40. 18 U.S.C. § 704.

41. *Alvarez*, 567 U.S. at 723.

42. *Id.* at 733 (Breyer, J., concurring).

43. *Id.* Note that in addition to a mens rea requirement, the Court requires actual harm to manifest from the false statement. The Court placing a requirement for mens rea with respect to false statements is a common occurrence. *See, e.g.*, Gertz v. Robert Welch, Inc., 418 U.S. 323, 375–76 (White, J., dissenting) (1974); Smith v. California, 361 U.S. 147, 150 (1959). The mens rea requirement and other standards that we impose on false statements can be ways in which the courts try to minimize the manifestation of the chilling effect. A more heightened intentional standard makes it less likely that speech will be chilled, while a lower (say, negligence) standard would make it more likely.

44. *See, e.g.*, Ashcroft v. Am. C.L. Union, 542 U.S. 656, 661 (2004); Reno v. Am. C.L. Union, 521 U.S. 844, 871–72 (1997); New York v. Ferber, 458 U.S. 747, 764–65, 772 (1982).

45. *See, e.g.*, Ashcroft v. Free Speech Coal., 535 U.S. 234, 234 (2002) (noting that the opponents of the broad text of the Child Pornography Prevention Act of 1996 (CPPA) alleged "that the 'appears to be' and 'conveys the impression' provisions" of the CPPA "are overbroad and vague, chilling production of works protected by the First Amendment").

46. *See, e.g.*, Illinois *ex rel.* Madigan v. Telemarketing Assocs., Inc., 538 U.S. 600, 620 (2003).

47. *See, e.g.*, Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 367–69 (2010); Cent. Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n, 447 U.S. 557, 564 n.6 (1980).

48. *See, e.g.*, Time, Inc. v. Hill, 385 U.S. 374, 386 (1967).

49. *See, e.g.*, Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 52–53, 56 (1988).

50. *See, e.g.*, Ariz. Free Enter. Club's Freedom Club PAC v. Bennett, 564 U.S. 721, 746 (2011). For various other cases implicating the chilling effect, see Virginia v. Hicks, 539 U.S. 113, 119 (2003); Sec'y of State of Md. v. Joseph H. Munson Co., 467 U.S. 947, 967–68 (1984); Laird v. Tatum, 408 U.S. 1, 11–13 (1972); United States v. Nat'l Treasury Emps. Union, 513 U.S. 454, 468 (1995) (discussing the First Amendment in the context of an honorarium ban).

51. *See generally, e.g.*, Kaminski, *supra* note 2; Kendrick, *supra* note 2; Schauer, *supra* note 2; Youn, *supra* note 1.

B. Recent Applications to Private Companies

While historically, the chilling effect has been limited to state actions or actions that at least have some form of state action in the causal chain,[52] more recently, the chilling effect has also broadened its scope to apply to private forms of speech restrictions as well.[53] Rather than courts doing this expansion, private actors have been the proponents of extending the doctrine outside the public sphere. This is ever present, especially when looking at the current activities, debates, and federal policies around social media platforms.

Facebook, Twitter, and other social media platforms are clearly implicated in the chilling effect phenomenon. These social media sites often restrict posts and activities on their platforms that are deemed fake news, political advertisements, and hate speech.[54] Social media sites like "Facebook and Twitter [have] terms and conditions [that] permit them to remove offending posts and photographs even without the consent of the individual who posted them. . . . Third-party intermediaries, therefore, frequently serve as censors."[55] Even the Trump administration argued that these restrictions are problematic as they chill speech, commenting in part:

---

52. *See* Youn, *supra* note 1, at 1496 ("As chilling effect doctrine developed, the category of consequences that were deemed to constitute an actionable chilling effect expanded to encompass a new and potentially problematic category: the consequences of private action.").

53. *Id.* at 1501–02 (characterizing this as "private chill"). *See also* John T. Bennett, *The Harm in Hate Speech: A Critique of the Empirical and Legal Bases of Hate Speech Regulation*, 43 HASTINGS CONST. L.Q. 445, 447 (2016) (arguing that the empirical evidence on the harms of hate speech are largely unexamined); Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1035–36, 1062 (2018) (arguing that the lack of clarity in speech restrictions can risk a type of global censorship creep); Matthew P. Hooker, *Censorship, Free Speech & Facebook: Applying the First Amendment to Social Media Platforms Via the Public Function Exception*, 15 WASH. J.L. TECH. & ARTS 36, 36 (2019) (discussing how social media platforms may implicate the First Amendment).

54. *See, e.g.*, Joel Timmer, *Fighting Falsity: Fake News, Facebook, and First Amendment*, 35 CARDOZO ARTS & ENT. L.J. 669, 699–703 (2017) (discussing the rise in fake news and Facebook's attempt to censor various political speech and the potential First Amendment implications thereof).

55. Jennifer M. Kinsley, *Chill*, 48 LOY. U. CHI. L.J. 253, 280 (2016) (citing *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards [https://perma.cc/3WAP-YSQ6] (containing Facebook's guidelines regarding when Facebook will remove posts and photographs)); *see also Twitter Terms of Service*, TWITTER, https://twitter.com/tos?lang-en [https://perma.cc/GKN2-8CBL] (noting the terms of service); *Programmable Search Engine Terms of Service*, GOOGLE, https://support.google.com/programmable-search/answer/1714300 [https://perma.cc/Y636-BUN2] (providing that "the Site [shall] not at any time contain any pornographic, hate-related, violent, or offensive content or contain any other material, products or services that violate or encourage conduct that would violate any criminal laws, any other applicable laws, any third party rights, or any service policies").

> In a country that has long cherished the freedom of expression, we cannot allow a limited number of online platforms to hand pick the speech that Americans may access and convey on the internet. This practice is fundamentally un-American and anti-democratic. When large, powerful social media companies censor opinions with which they disagree, they exercise a dangerous power.[56]

For example, take Facebook's policy of restricting nudity in posts. This policy could suppress valuable speech that takes the form of visual art or sexual education.[57] In addition, Facebook's "judgments about what content is gratuitously violent or hateful toward a religious or ethnic group can vary widely, and the result will be subjective and unpredictable censorship of literature, art, and political discussion."[58] As such, Facebook's content moderation decisions likely create a chilling effect, much like public regulations that censor speech are purposed to do.[59]

In January 2021, Facebook and Twitter suspended the account of President Donald Trump, with Twitter citing that many of his posts contained lies and incited violence.[60] In addition, the same platforms restricted posts that advocated for various political conspiracies that may have led to the Capitol protest on January 6, 2021.[61] This action, not surprisingly, further brought social media speech restrictions and

---

56. Preventing Online Censorship, 85 Fed. Reg. 34,069 § 1 (May 28, 2020).

57. Heins, *supra* note 6, at 326. *See generally* Marvin Ammori, *The "New"* New York Times: *Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259 (2014) (arguing that some of the most important speech regulations are happening at tech companies).

58. Heins, *supra* note 6, at 326.

59. *See, e.g.*, Jeff Kosseff, *First Amendment Protection for Online Platforms*, COMPUT. L. & SEC. REV., Oct. 2019, at 16 ("Exposing platforms to additional immunity could cause them to change business models and limit users' ability to share content online, possibly chilling legal and legitimate speech. The concerns about spillover effects on legitimate speech are very real."). In addition, these censorships have been criticized as helping to filter bubbles, which in turn may undermine democratic principles. *See* Philip M. Napoli, *What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble*, 70 FED. COMMC'NS. L.J. 55, 88 (2018) (arguing that social media censorship may create a market failure in the marketplace of ideas).

60. Mike Isaac & Kate Conger, *Facebook Bars Trump Through End of His Term*, N.Y. TIMES (Jan. 7, 2021), https://www.nytimes.com/2021/01/07/technology/facebook-trump-ban.html [https://perma.cc/6G4F-67PX]; *Permanent Suspension of @realDonaldTrump*, *supra* note 5.

61. *See, e.g.*, Haley Messenger, *Facebook Bans All 'Stop the Steal' Content*, NBC NEWS (Jan. 11, 2021, 5:28 PM), https://www.nbcnews.com/tech/tech-news/facebook-bans-all-stop-steal-content-n1253809 [https://perma.cc/948C-TFBH]; Brakkton Booker, *Facebook Removes 'Stop the Steal' Content; Twitter Suspends QAnon Accounts*, NPR (Jan. 12, 2021, 12:54 PM), https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/12/956003580/facebook-removes-stop-the-steal-content-twitter-suspends-qanon-accounts [https://perma.cc/79RG-5ZM6].

the chilling effect into the spotlight. Legislatures have threatened to introduce legislation to prohibit the ability of social media platforms to restrict posts based upon their content.[62]

One might wonder why these platforms are not violating the First Amendment given that the censorship they create has the same kind of chilling effect that public regulations do. Sanctions like flagging one's account and getting banned from the platform seemingly create serious chilling effects, not to mention potential reputational harm.[63] Users may be deterred from their actions to avoid engaging in any type of speech or related content that is banned on the platforms. This chilling is created in part due to the uncertain and broad terms of service that social media platforms utilize. The effect is likely further strengthened by the lack of clarity on how certain sanctions are applied and adjudicated.[64] Scholars explicitly recognize that private companies engage in actions that chill speech in similar ways as public regulations do, discussing the phenomenon as legal private chill.[65]

Most simply, these companies are private and hence not subject to the government limitations presented in the First Amendment.[66] But the thorniness of private speech censorship goes even further. Specific federal regulations have been passed that protect social media sites from litigation relating to their decisions to censor speech and the content of what they do publish. 47 U.S.C. § 230 ("Section 230"), as part of the Communications Decency Act, immunizes social media companies from liability associated with any third-party content they publish.[67]

---

62. Former President Trump sought to narrow Section 230 immunity, which allows companies like Twitter and Facebook to restrict content. Bobby Allyn, *Stung by Twitter, Trump Signs Executive Order to Weaken Social Media Companies*, NPR (May 28, 2020, 4:59 PM), https://www.npr.org/2020/05/28/863932758/stung-by-twitter-trump-signs-executive-order-to-weaken-social-media-companies [https://perma.cc/H5PG-C39A].

63. *See, e.g.*, Elizabeth Dwoskin, Nitasha Tiku & Heather Kelly, *Facebook to Start Policing Anti-Black Hate Speech More Aggressively than Anti-White Comments, Documents Show*, WASH. POST (Dec. 3, 2020, 8:00 AM), https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/ [https://perma.cc/RUS6-KNCB].

64. *See* Heins, *supra* note 6, at 325–26 ("But there is no judicial determination of illegality — just the best guess of Facebook's censors. Facebook's internal appeals process is mysterious at best.").

65. Youn, *supra* note 1, at 1476–81. Youn creates three categories of chilling: government chill, illegal private chill, and legal private chill. *Id.* at 1537. Youn focuses her arguments on which forms of private speech chilling are inappropriate. *See id.* at 1510–20. For the purposes of this Article, it does not matter which forms are illegal and whether they implicate the First Amendment. This Article is solely focused on whether speech actually gets chilled in these contexts and how we can measure it in order to better apply it in various legal and non-legal cases.

66. Of course, there are arguments that social media platforms should be treated more like public entities with respect to the state action doctrine. Hooker, *supra* note 53, at 38.

67. For a general overview of Section 230 immunity, see generally Eric Goldman, *An Overview of the United States' Section 230 Internet Immunity*, *in* THE OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY (Giancarlo Frosio ed., 2020). *See also* Eric Goldman, *Why Section 230 Is Better than the First Amendment*, 95 NOTRE DAME L. REV. REFLECTION 33, 34 (2019) (supporting Section 230 and arguing that it has more protections than the First

This regulation aims to prevent private social media platforms from being held liable for defamation and invasion of privacy. By allowing companies to set their own policies, some have argued that Section 230 allows social media platforms to refrain from engaging in private censorship which in turn can promote free speech on the Internet.[68]

While Section 230 encourages free speech, it also does not explicitly prohibit private restriction,[69] indeed it "affirmatively allows it, and therein lies the rub: [Section 230 allows] vague, broad terms of service, applied by powerful companies like Facebook with no transparency and no clear avenues for appeal."[70] In effect, given the protections granted by federal law, private companies are the "arbiters of what gets communicated in the brave new world of cyberspace."[71]

The opportunities social media companies have to legally restrict speech such as fake news, political speech, hate speech, and other profanity on their platforms have created a normative debate around whether social media companies should limit certain types of speech. Some have argued that it is beneficial and prevents harms associated with fake news, conspiracies, hate speech, etc. For example, Philip Napoli argues that, in theory, counterspeech is good for the marketplace of ideas.[72] Allowing fake news and real news to battle it out is what the First Amendment is all about. However, he notes that technological changes call into question whether counterspeech is effectively making the marketplace of ideas efficient.[73] As such, it follows that certain types of private censorship may actually be necessary in order to disseminate real (as opposed to fake) news. Others have argued that hate

---

Amendment does and hence should not be curtailed); Eric Goldman, *The Complicated Story of FOSTA and Section 230*, 17 FIRST AMENDMENT L. REV. 279, 288 (2019) (discussing the relationship between legislation outlawing sex trafficking and Section 230 immunity).

68. Kosseff, *supra* note 59, at 1.

69. "No provider or user of an interactive computer service shall be held liable on account of . . . any action voluntarily taken in good faith to restrict access to or avail- ability of material that the provider or user considers to be obscene, lewd . . . or otherwise objectionable . . . ." 47 U.S.C. § 230(c)(2) (2012).

70. Heins, *supra* note 6, at 328.

71. *Id.* at 325.

72. *See* Napoli, *supra* note 59, at 88–90.

73. *See id.* at 87. The marketplace of ideas is a common analogy first articulated by Justice Holmes in *Abrams v. United States*, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting). He argued that the best way for truth to disseminate in public discourse is for ideas to be freely traded in society. The best ideas are promoted and the worst ideas are demoted. For a more detailed description of the marketplace of ideas, see generally R. H. Coase, *The Market for Goods and the Market for Ideas*, 64 AM. ECON. REV. 384 (1974). For a discussion of the concept of the marketplace of ideas, see generally Stanley Ingber, *The Marketplace of Ideas: A Legitimizing Myth*, 1984 DUKE L.J. 1 (1984); Maxwell E. McCombs & Donald L. Shaw, *The Evolution of Agenda-Setting Research: Twenty-Five Years in the Marketplace of Ideas*, J. COMM., June 1993, at 58, 58; Vincent Blasi, *Holmes and the Marketplace of Ideas*, SUP. CT. REV., 2004, at 1.

speech causes psychological harm and hence has no place in online posts.[74]

In contrast, there have been several arguments criticizing private speech restrictions on the internet as they violate free speech principles and hinder the efficiency of the marketplace of ideas. These scholars argue that the restricting hate speech can create chilling effects.[75] Some have argued that the negative effects of hate speech rely upon empirical harms that have gone undocumented and therefore the speech does not need to be restricted.[76] Still others have argued that the best way to combat hate speech is with more speech, not censorship.[77]

Given the recent and likely continuing debate around the chilling effect in private speech restrictions, it is incredibly important that empirical work begins to grapple with whether speech is actually chilled. Determining whether and how speech actually is chilled will not only help legal scholars and judges better understand which regulations will have chilling effects but will also give decision makers in private companies more confidence in their content restriction decisions. As such, measuring the chilling effect is an incredibly important endeavor.

C. Existing Empirical Work

Most studies that have attempted to measure the chilling effect have focused on surveys and used the instance of defamation and libel laws to assess whether respondents' speech is chilled.[78] These studies usually ask respondents how likely they are to be affected by a specific regulation. For example, a study may poll reporters or editors of news

---

74. *See* Stefanie Ullmann & Marcus Tomalin, *Quarantining Online Hate Speech: Technical and Ethical Perspectives*, 22 ETHICS & INFO. TECH. 69, 71 (2020) ("[I]n response to growing public concerns about [hate speech], most social media platforms have adopted self-imposed definitions, guidelines, and policies for dealing with this particular kind of offensive language.").

75. Ronald Dworkin, Foreword to EXTREME SPEECH AND DEMOCRACY, v-viii (Ivan Hare & James Weinstein eds. 2009); James Weinstein, *Hate Speech Bans, Democracy, and Political Legitimacy*, 32 CONST. COMMENT. 527, 527 (2017); Ullmann & Tomalin, *supra* note 74, at 74 ("The various disagreements have centered on topics such as whether [hate speech] bans necessarily undermine democratic legitimacy by depriving certain citizens of a voice in the political process . . . . Contrasting views about such matters become vividly apparent in relation to online [hate speech] . . . .").

76. *See generally* John T. Bennett, *The Harm in Hate Speech: A Critique of the Empirical and Legal Bases of Hate Speech Regulation*, 43 HASTINGS CONST. L.Q. 445 (2016).

77. *See generally* NADINE STROSSEN, HATE: WHY WE SHOULD RESIST IT WITH FREE SPEECH, NOT CENSORSHIP (2018).

78. *See generally* ERIC BARENDT, LAURENCE LUSTGARTEN, KENNETH NORRIE & HUGH STEPHENSON, LIBEL AND THE MEDIA: THE CHILLING EFFECT (1997) (using interviews to analyze the chilling effect of English libel rules); David A. Logan, *Libel Law in the Trenches: Reflections on Current Data on Libel Litigation*, 87 VA. L. REV. 503, 511 (2001) (finding that chilling does not really have a large effect in America); Chris Dent & Andrew T. Kenyon, *Defamation Law's Chilling Effect: A Comparative Content Analysis of Australian and US Newspapers*, 9 MEDIA & ARTS L. REV. 89 (2004).

outlets and ask them to rate how likely they would be to not publish something given a certain censorship regime.[79]

In one study, Jonathon Penney asked respondents on a scale how likely they were to "speak or write about certain topics online" in various hypothetical scenarios that implicated speech censorship.[80] In that same study, Penney also measured whether respondents would be more careful in what they said online in the same hypothetical speech censorship scenarios.[81] His results showed that when asked hypothetically if a given censorship regulation would chill their speech, respondents overwhelmingly said that it would. Similar studies have yielded similar results.[82] This is partly due to the substantial amount of self-reporting bias that exists in these kinds of survey studies.[83] Respondents think that they will be chilled or like to portray a type of identity in these kinds of studies, and this creates a bias in their responses. In other cases, there have been different effects, which should not be surprising because these studies use different samples (lawyers, editors, news writers, etc.), different levels of censorship, and different forms of sanctions.[84]

---

79. *See, e.g.*, BARENDT ET AL., *supra* note 78 (conducting interviews with various individuals); Stephen M. Renas, Charles J. Hartmann & James L. Walker, *An Empirical Analysis of the Chilling Effect: Are Newspapers Affected by Liability Standards in Defamation Actions?*, *in* THE COST OF LIBEL: ECONOMIC AND POLICY IMPLICATIONS (Everette E. Dennis & Eli M. Noam eds., 1989) (surveying editors of U.S. newspapers); Michael Massing, *The Libel Chill: How Cold* Is *It Out There?*, 24 COLUM. JOURNALISM REV. 31 (1985) (interviewing attorneys and news editors).

80. Penney, *supra* note 11, at 7. In that study, Penney used four different hypothetical situations including an anti-cyberbullying statute, government surveillance, private surveillance, and personal legal threats. Penney used a Likert Scale to measure his dependent variable. A Likert Scale is a common way to measure how likely a respondent is to take a certain action. There are many forms of the scale, and the scale often runs from either 1 to 5 or 1 to 7. For a detailed discussion of the Likert Scale, see Ankur Joshi, Saket Kale, Satish Chandel & D. K. Pal, *Likert Scale: Explored and Explained*, 7 BRIT. J. APPLIED SCI. & TECH. 4 (2015).

81. Penney also asked several other questions, which all took the form of "how likely" one would be to do something given the censorship. Penney, *supra* note 11, at 5, 7–8.

82. *See, e.g.*, Weaver & Bennett, *supra* note 28, at 1189 (concluding that interview subjects did admit to experiencing a chilling effect); David A. Barrett, *Declaratory Judgments for Libel: A Better Alternative*, 74 CALIF. L. REV. 847, 860 (1986) (concluding that, apart from large media defendants, many speakers experience the chilling effect); Kendrick, *supra* note 2, at 1678.

83. Self-reporting bias is a common occurrence in this kind of empirical work but has even been noted in the work on chilling. *See, e.g.*, Kendrick, *supra* note 2, at 1679 ("[Interviews] have drawbacks . . . . For instance, media participants may be unwilling to admit that they sacrificed journalistic principles out of fear of litigation, or they may be willing to exaggerate the chilling effect of the law in order to downplay other considerations that informed a decision to kill or revise a story.") and accompanying text.

84. *See, e.g.*, Renas et al., *supra* note 79 (using four hypothetical manipulated legal rules to see how editors will react); Jeremy Cohen, Diana Mutz, Vincent Price & Albert Gunther, *Perceived Impact of Defamation: An Experiment on Third-Person Effects*, 52 PUB. OP. Q. 161, 161 (1988) (using an experimental approach to ask whether others thought people would be chilled).

The problem with these studies, however, is that they are only seeking to understand whether individuals would claim that they would be affected by a certain censorship regime. No study has actually imposed a specific instance of censorship and measured how people behaved. To determine whether speech is chilled, one needs to collect data on some output of speech and then analyze how or whether the output changes in the face of actual, not hypothetical, censorship. Existing studies are limited because simply asking somebody whether they would hypothetically change their speech is not the most rigorous or reliable way to measure manifestations of a chilling effect.

In addition, these studies only measure the chilling effect in one form. That is, these studies are focused on whether a respondent would engage in protected speech when facing censorship. As described above, the chilling effect implicates more than this conception of deterrence. Individuals may still speak but may change their speech in subtle or significant ways, which would still evidence a form of the chilling effect.[85] Consequently, to measure whether the chilling effect exists and in what ways speech is chilled, a study would need to analyze actual speech for subtle or significant differences, not just ask respondents what they would hypothetically do or not do.

The difficulty of empirically measuring the chilling effect is not underestimated — many have noted that truly empirically measuring the chilling effect is quite challenging.[86] The difficulty exists because one needs a baseline to measure how speech changes. In other kinds of empirical legal studies, scholars have used a change in the law to compare behavior pre- and post-implementation of the law. In that way, the behavior pre-implementation can act as a baseline to which behavior post-implementation can be compared.[87] One such study compared pre- and post-implementation abortion rates in the face of changing abortion laws and showed that abortion rates were chilled.[88]

To measure the chilling effect before and after a legal change, also called an event study, we would need to anticipate a law or court ruling that would substantially limit speech. Then, we would compare the

---

85. *See supra* note 28 and accompanying text.

86. *See, e.g.*, Schauer, *supra* note 2, at 730 (noting that many measures of the chilling effect, such as predictions of human behavior, are most likely unprovable empirically); Kendrick, *supra* note 2, at 1675 ("It is difficult to establish either the presence or the absence of a chilling effect, let alone to measure the extent of such an effect.").

87. This kind of study is common in empirical law and economic studies. *See generally* Sanjai Bhagat & Roberta Romano, *Event Studies and the Law: Part I: Technique and Corporate Litigation*, 4 AM. L. & ECON. REV. 141 (2002) (discussing the general approach to using event studies in legal research); Sanjai Bhagat & Roberta Romano, *Event Studies and the Law: Part II: Empirical Studies of Corporate Law*, 4 AM. L. & ECON. REV. 380 (2002) (focusing just on the use of event studies in corporate law).

88. Canes-Wrone & Dorf, *supra* note 11, at 1097–98 (using an event study to show that changes in policy affecting abortion timelines did have a chilling effect). This was not speech, but rather, as the authors recognize, a study of chilling of behavior.

speech of citizens before and after the law or court ruling went into effect to see if there is a difference. We would also need to make the challenging decisions of which citizens to study, which aspects of their speech to examine, and how to measure the differences. But before one even gets to that point, finding such a law and anticipating it at the right moment, so as to effectively collect data, is all but impossible.

Below, this Article attempts to address the difficulties and weaknesses of previous studies which have measured the chilling focusing only on hypothetical chilling. This study, instead, experimentally manipulates censorship and then compares the speech output of respondents in various manipulated conditions. This is a more controlled and approachable way to mimic a real change of law.[89]

III. MEASURING THE CHILLING EFFECT

A. Overview of the Study

The empirical strategy of this Article is to use an experimental manipulation to measure how speech changes in the face of restrictions. Using three conditions — one with no censorship, one with specific censorship, and one with broad censorship — the study asked respondents to write Google/Yelp-like reviews of a recent dining experience. Importantly, the respondents were directed to focus only on an experience that was negative and to write an extremely negative review.

One group of respondents was asked to write negative reviews without any restrictions. This group served as a baseline for uncensored speech. A second group was asked to perform the same task but was specifically told not to use certain words. The prohibited words did not appear in the reviews of the baseline condition. A third group was asked to perform the same task but specifically told not to use "hate speech, profanity, or offensive language." This was a broad restriction intended to chill speech more than the second condition. Notably, no respondents in the baseline condition used anything that would have risen to the level of "hate speech, profanity, or offensive language."

Text analysis was then used to analyze each group's negative reviews. The results indicate that there is a statistically significant difference in the texts of the experimental groups. The group that created the most positively toned reviews (as measured by the number of words with a positive sentiment) was the broad censorship group, and the group that created the least overall positively toned reviews (as measured by the least number of words with a positive sentiment) was the no censorship group. However, notably, the negativity as measured by

89. There are of course weaknesses to an experimentally manipulated type of study. Those are discussed further in Part IV.

the number of negative words of each group's reviews did not change. In addition, further testing showed that third parties were equally persuaded by each group's reviews (i.e., the groups ended up sending the same message). That is, although the censored groups changed their speech, and hence were chilled in the strict sense of the effect, the substance and communicative effect of the speech did not change.

B. Design of the Study

The study used online restaurant reviews as a type of speech. In particular, the study asked respondents to write negative online reviews of a dining experience they had in the past six months. Of course, there are other forms of speech that are arguably more valuable in terms of the First Amendment. For example, political speech or speech in an academic setting may implicate more the importance of the First Amendment because these are historically instances where free speech is deemed to be incredibly valuable. But online restaurant reviews have some characteristics that make them useful for a study like this. First, online reviews are something that many people around the world actively partake in. Speech like political commentary or news criticism is not necessarily universally understood or practiced to the same extent. But almost everybody has either interacted with or written an online review.[90] Therefore, using reviews for the study provided a universal form of speech that everybody in the sample would understand.

Second, online reviews have been used in many empirical contexts. A significant amount of marketing and consumer psychology research has utilized respondents writing online reviews to better understand how consumers interact with online content, online providers, social media, etc.[91] Third, given that social media censorship and online

90. *See* Diana Kaemingk, *Online Reviews Statistics to Know in 2021*, QUALTRICS (Oct. 30, 2020), https://www.qualtrics.com/blog/online-review-stats/ [https://perma.cc/63PH-VNAQ].

91. *See generally, e.g.*, Eric K. Clemons, Guodong Gordon Gao & Lorin M. Hitt, *When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry*, 23 J. MGMT. INFO. SYS. 149 (2006) (analyzing online reviews for beers to show that variance of ratings affect product growth); Wenjing Duan, Bin Gu & Andrew B. Whinston, *Do Online Reviews Matter? — An Empirical Investigation of Panel Data*, 45 DECISION SUPPORT SYS. 1007 (2008) (analyzing online reviews in the context of movies); Jonah Berger, Alan T. Sorensen & Scott J. Rasmussen, *Positive Effects of Negative Publicity: When Negative Reviews Increase Sales*, 29 MKTG. SCI. 815 (2010) (showing how negative online reviews can actually benefit companies); Ann E. Schlosser, *Can Including Pros and Cons Increase the Helpfulness and Persuasiveness of Online Reviews? The Interactive Effects of Ratings and Arguments*, 21 J. CONSUMER PSYCH. 226 (2011) (showing that presenting one versus two sides in an online review is often more persuasive); Philip Fei Wu, *In Search of Negativity Bias: An Empirical Study of Perceived Helpfulness of Online Reviews*, 30 PSYCH. & MKTG. 971 (concluding that negative reviews are no more helpful than positive reviews on Amazon.com); Fahri Karakaya & Nora Ganim Barnes, *Impact of Online Reviews of Customer Care Experience on Brand or Company Selection*, 27 J. CONSUMER MKTG. 447 (2010) (finding that consumer reviews are more important to purchasing decisions than government-sponsored information sites and company websites).

speech censorship writ large are incredibly important topics considering current events,[92] focusing on a typical online form of speech provides an immediately relevant case study.

To observe if the speech in online reviews was chilled, the study utilized private censorship. In particular, the study experimentally manipulated three conditions. The first condition (the no censorship condition) did not impose any kind of censorship on the online negative reviews that respondents were asked to write. This condition was meant to replicate an existing online review platform (e.g., Yelp or Google Reviews). The second condition (the specific censorship condition) imposed censorship that focused on specific words. In the specific censorship condition, respondents were told that they were not allowed to use various words in their online negative reviews. Those words were: "repugnant, putrid, gruesome, vomit, vile, noxious, revolting, sh*t, f*ck, a**hole."

The words in the specific censorship condition were chosen after a pilot test using the no censorship condition so that none of the words appeared in the pilot test reviews.[93] This is an important point. The censorship that was imposed was designed to not actually restrict the online reviews in a substantial way. Given that we can assume statistically that the reviews in the pilot test were representative of reviews in any given sample due to random selection, the study chose words to censor that did not appear in the pilot test reviews.

By censoring only words unlikely to occur in the reviews themselves, the study created a censorship condition that, absent a chilling effect, should not affect the online reviews. If a respondent was never going to use any of the forbidden words in the specific censorship condition and was not chilled by the existence of the censorship, the text of their online review should not change. If, however, the text of the

92. *See supra* Section II.B. Social media has also been of prime importance in business scholarship and consumer psychology scholarship. *See generally, e.g.*, Olivier Toubia & Andrew T. Stephen, *Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter?*, 32 MKTG. SCI. 368 (2013) (discussing motivations for contributing to social media); Andrew T. Stephen, *The Role of Digital and Social Media Marketing in Consumer Behavior*, 10 CURRENT OP. PSYCH. 17 (2016) (analyzing how information from social media affects purchasing decisions); THE DARK SIDE OF SOCIAL MEDIA: A CONSUMER PSYCHOLOGY PERSPECTIVE (Angeline Close Scheinbaum ed., 2018) (arguing that social media can also have harmful effects); Vikram R. Bhargava & Manuel Velasquez, *Ethics of the Attention Economy: The Problem of Social Media Addiction*, 31 BUS. ETHICS Q. 321 (2021) (arguing that social media platforms have ethical obligations to curtail consumer addiction).

93. Pilot testing in consumer psychology studies is very common. It allows a researcher to better understand how respondents will react to various conditions, survey designs, and prompts without having to spend a lot of money or exhaust large samples. *See generally* Jason M. Etchegaray & Wayne G. Fischer, *Understanding Evidence-Based Research Methods: Pilot Testing Surveys*, 4 HEALTH ENV'TS RSCH. & DESIGN J. 143 (2011); Norbert Schwarz, Robert M. Groves & Howard Schuman, *Survey Methods*, *in* HANDBOOK OF SOCIAL PSYCHOLOGY (4th ed. 1998).

review is different when imposing the specific censorship, then this is evidence that the censorship created a chilling effect.

The third condition (the broad censorship condition) imposed censorship that focused on general categories of words or phrases. Specifically, respondents were told that they were not allowed to use the following types of words: "NO Foul language, NO Racially charged language, NO Offensive language." Again, the categories in this condition were chosen in the same way the words in the second condition were chosen — making sure that none of the categories of words had appeared in the pilot test. Thus, like the second condition, the broad censorship condition was designed not to reach the content of the pilot reviews written under the baseline conditions. If the text of the review was different in the broad censorship condition, then this would be evidence that the censorship created a chilling effect.

These three conditions also allowed the study to tease out differences in chilling effects in the context of different forms of censorship. The chilling effect is predicted to occur because individuals are uncertain as to what exact speech is censored. In that way, the broad censorship condition should show more chilling than the specific censorship condition.[94] And if uncertainty is really a necessary condition for chilling, we would predict that the specific censorship condition would not show any chilling whatsoever given that there is no uncertainty.

An empirical test of public government-oriented censorship would be ideal. Given that this is private censorship, one could argue that the results described may not fully translate to a public form of censorship. However, public government-oriented censorship is difficult to manipulate reliably and convincingly through online surveys.[95] As such, this study used a private censorship that mimicked a type of believable social media speech restriction. However, understanding private censorship chilling is important both because it is currently purported to

94. One may respond that the broad censorship condition does not create uncertainty given that based upon the pilot tests, respondents did not include words that would fall into the broad censorship categories. Although this is true, the broad condition creates more uncertainty than the specific conditions. So, although it may not create much uncertainty overall, comparatively it should create more uncertainty than the other conditions.

95. Testing public government-oriented censorship is difficult for various reasons. First, it does not happen a lot. Ideally, we would want an event study with several forms of a speech censorship law that were imposed on differing populations. This is grossly impractical and would cost several hundreds of thousands of dollars and require several hundred government officials and policy makers to effectuate. As such, a simple experimental fabricated public regulation would be second best. However, to make this believable, there would need to be some kind of government sanction that respondents would face if they ignored the censorship regulation. A private researcher cannot impose a government sanction and thus any type of government-used sanction would only be hypothetical. Previous studies have adopted such hypothetical sanctions (*see* Renas et al., *supra* note 79; Cohen et al., *supra* note 84). While useful, this Article goes beyond the hypothetical and actually imposes a real censorship restriction to see how individuals behave.

happen and because it still gives us insight into the existence of public censorship chilling.

According to the theory of the chilling effect, there needs to be a sanction that occurs if an individual violates a given censorship rule. Therefore, to see the chilling effect in this study, some type of sanction would need to be placed on respondents if they used words or groups of words that fell into the censorship conditions. Given that this study was run online via Amazon Mechanical Turk,[96] the simple sanction was to withhold the compensation that respondents received from going through the study if they violated a censorship condition.

The respondents were paid $2 for their time,[97] but the payment only occurred after a respondent completed the full study. If a respondent did not follow directions, missed an attention check, or otherwise left the study, they did not receive their payment. Respondents were informed of the censorship conditions and the sanction: if their online reviews contained words or categories of words that were censored, they would not be paid the $2 that each respondent would receive if they followed directions.[98] Given this sanction and the potential for uncertainty on what counted as a censored word, the design of the study matched the uncertain environment necessary for a chilling effect.

In order to analyze the online reviews (i.e., the speech), the study utilized text analysis. Text analysis is a relatively new way to analyze the content of written or vocalized text.[99] It has been used in various

---

96. Amazon Mechanical Turk is an online marketplace that allows businesses and individuals to quickly coordinate with human subjects to perform tasks. This includes fielding surveys and other empirical studies for many social scientists.

97. This is consistent with the going rate on Mechanical Turk. For a discussion of pay rates on Amazon Mechanical Turk, see generally Michael Buhrmester, Tracy Kwang & Samuel D. Gosling, *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?*, *in* METHODOLOGICAL ISSUES AND STRATEGIES IN CLINICAL RESEARCH 133–39 (Alan E. Kazdin ed., 2016).

98. This kind of sanction is commonly used in various marketing and consumer psychology research, in particular research that is focused on online content creation. *See generally* Nicolas Kaufmann, Thimo Schulze & Daniel Veit, Conference Paper, *More Than Fun and Money. Worker Motivation in Crowdsourcing — A Study on Mechanical Turk*, 17 AMS. CONF. ON INFO. SYS., Jan. 2011, at 1 (discussing how motivation theory predicts use of Mechanical Turk); David G. Rand, *The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments*, 299 J. THEORETICAL BIO. 172 (2011) (discussing how Mechanical Turk can be used with sanctions to study various behaviors). In addition, many researchers use bonuses and sanctions in the context of dictator (or other cooperative) based game studies. *See, e.g.*, Nichola J. Raihani, Ruth Mace & Shakti Lamba, *The Effect of $1, $5 and $10 Stakes in an Online Dictator Game*, 8 PLOS ONE e73131 (2013); Pablo Brañas-Garza, Valerio Capraro & Ericka Rascón-Ramírez, *Gender Differences in Altruism on Mechanical Turk: Expectations and Actual Behaviour*, 170 ECON. LETTERS 19 (2018); Ofra Amir, David G. Rand & Ya'akov Kobi Gal, *Economic Games on the Internet: The Effect of $1 Stakes*, 7 PLOS ONE e31461 (2012); Kyle A. Thomas & Scott Clifford, *Validity and Mechanical Turk: An Assessment of Exclusion Methods and Interactive Experiments*, 77 COMPUTS. HUM. BEHAV. 184 (2017).

99. *See* Mary C. Lacity & Marius A. Janson, *Understanding Qualitative Data: A Framework of Text Analysis Methods*, 11 J. MGMT. INFO. SYS. 137, 139 (1994). For a detailed

contexts ranging from political science,[100] marketing/consumer psychology,[101] and even sociology.[102] However, text analysis has been surprisingly absent in legal research.[103] This method can be useful for legal scholarship not just for court cases and opinions, but also studies focusing on speech and the chilling effect.[104]

This particular study used the text analysis program LIWC 2015 (Linguistic Inquiry and Word Count).[105] In effect, text analysis examines all the words in a given set of sentences or paragraphs and then characterizes and scores them. LIWC does this using its own extensive dictionary that labels each word on various dimensions including linguistic dimensions (e.g., is it a noun, verb, adjective, etc.?) and psychological dimensions (e.g., does the word elicit anger, sadness, happiness, etc.?). The program then uses these labels to score words. It then combines the score of words for any given string of text to score a whole

---

background on content/text analysis, see KLAUS KRIPPENDORFF, CONTENT ANALYSIS: AN INTRODUCTION TO ITS METHODOLOGY (2d ed. 2004).

100. *See, e.g.*, Will Lowe, *Understanding Wordscores*, 16 POL. ANALYSIS 356 (2008) (detailing the method of wordscores, a form of text analysis); Burt L. Monroe & Philip A. Schrodt, *Introduction to the Special Issue: The Statistical Analysis of Political Text*, 16 POL. ANALYSIS 351 (2008) (discussing how text analysis can be used to analyze political texts); Kenneth Benoit & Michael Laver, *Estimating Irish Party Policy Positions Using Computer Wordscoring: The 2002 Election — A Research Note*, 18 IRISH POL. STUD. 97 (2003) (using text analysis to analyze Irish politics); Justin Grimmer & Brandon M. Stewart, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, 21 POL. ANALYSIS 267 (2013) (discussing how text analysis can be used in political science research).

101. *See, e.g.*, Jonah Berger, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer & David A. Schweidel, *Uniting the Tribes: Using Text for Marketing Insight*, 84 J. MKTG. 1 (2020) (detailing the use of text analysis in marketing research); Seshadri Tirunillai & Gerard J. Tellis, *Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation*, 51 J. MKTG. RSCH. 463 (2014) (using text analysis and big data to glean latent characteristics of consumer behavior); Praveen Aggarwal, Rajiv Vaidyanathan & Alladi Venkatesh, *Using Lexical Semantic Analysis to Derive Online Brand Positions: An Application to Retail Marketing Research*, 85 J. RETAILING 145 (2009) (tracking brand position using text analysis).

102. *See, e.g.*, Jeremiah Bohr & Riley E. Dunlap, *Key Topics in Environmental Sociology, 1990–2014: Results from a Computational Text Analysis*, 4 ENV'T. SOCIO. 181 (2017) (introducing sociologists to content analysis methods).

103. Some notable exceptions include: Mark A. Hall & Ronald F. Wright, *Systematic Content Analysis of Judicial Opinions*, 96 CALIF. L. REV. 63 (2008); Chad M. Oldfather, Joseph P. Bockhorst & Brian P. Dimmer, *Triangulating Judicial Responsiveness: Automated Content Analysis, Judicial Opinions, and the Methodology of Legal Scholarship*, 64 FLA. L. REV. 1189 (2012); Michael Evans, Wayne McIntosh, Jimmy Lin & Cynthia Cates, *Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research*, J. EMPIRICAL LEGAL STUD. 1007 (2007).

104. One goal of this Article is also to introduce text analysis more broadly to the legal research in the hopes that it will spur more empirical scholarship.

105. The program can be found at LIWC, http://liwc.wpengine.com [https://perma.cc/BV2V-98R4]. Included there are a dictionary and detailed specifications on how the program works and was created. In addition, the following document spells out the exact characteristics LIWC uses to score text. JAMES W. PENNEBAKER, RYAN L. BOYD, KAYLA JORDAN & KATE BLACKBURN, THE DEVELOPMENT AND PSYCHOMETRIC PROPERTIES OF LIWC2015 (2015), https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf [https://perma.cc/M7BV-LWNT].

series of sentences on both linguistic and psychological dimensions. For example, the program may detect several psychologically positive words in a string (e.g., good, great, best, happy) or psychologically negative words (e.g., bad, worst, terrible) and then give a total positive and negative score for a given sentence or set of sentences. Doing this allows for the program to compare sets of text and conclude that one has more negative words or negative sentiment than another.

After analyzing the text via LIWC 2015, the study selected a random sample of reviews from each of the conditions and had third parties rate each of the reviews. This analysis helped determine to what extent the full communicative message of each condition's reviews changed or stayed the same.

1. Sample

The study used Amazon Mechanical Turk in combination with CloudResearch[106] to recruit respondents to partake in the study. Amazon Mechanical Turk is an online marketplace that allows businesses and individuals to quickly coordinate with human subjects to perform tasks. This includes fielding surveys and other empirical studies for many social scientists. Thousands of articles from disciplines including psychology, sociology, marketing, management, political science, and the law have utilized Mechanical Turk samples.[107] Mechanical Turk respondents have been shown to be just as reliable as laboratory experiments in most cases.[108] Using this online marketplace to produce reliable and valid results has become a norm in social science.[109]

106. Leib Litman, Jonathan Robinson & Tzvi Abberbock, *TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences*, 49 BEHAV. RES. METHODS 433 (2016) (discussing TurkPrime, which is now known as CloudResearch).

107. Thousands of articles have used, and continue to use, Amazon Mechanical Turk. The following is a non-exhaustive list of articles that used the online marketplace: Thomas Stevens, Aaron K. Hoshide & Francis A. Drummond, *Willingness to Pay for Native Pollination Of Blueberries: A Conjoint Analysi*s, 2 INT'L J. AGRIC. MKTG. 68 (2015); Karoline Mortensen & Taylor L. Hughes, *Comparing Amazon's Mechanical Turk Platform to Conventional Data Collection Methods in the Health and Medical Research Literature*, 33 J. GEN. INTERNAL MED. 4 (2018); Kirk Bansak, Jens Hainmueller, Daniel J. Hopkins & Teppei Yamamoto, *The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments*, 26 POL. ANALYSIS 112 (2018); Cindy Wu et al., *What Do Our Patients Truly Want? Conjoint Analysis of an Aesthetic Plastic Surgery Practice Using Internet Crowdsourcing*, 37 AESTHETIC SURGERY J. 105 (2017); Yu Pu & Jens Grossklags, *Using Conjoint Analysis to Investigate the Value of Interdependent Privacy in Social App Adoption Scenarios*, 36 INT'L CONF. ON INFO. SYS. (2015).

108. *See* Buhrmester et al., *supra* note 97 (arguing that Amazon Mechanical Turk respondents are more diverse and the data obtained is just as reliable as more traditional methods); Frank R. Bentley, Nediyana Daskalova & Brook White, *Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys*, 2017 CHI CONF. EXTENDED ABSTRACTS ON HUM. FACTORS COMPUT. SYS. 1092 (discussing the reliability of traditional marketplace consumer research versus Amazon Mechanical Turk).

109. *See* Buhrmester et al., *supra* note 97; Bentley et al., *supra* note 108.

CloudResearch, formerly called TurkPrime, is an independent company that allows researchers to recruit panels from Mechanical Turk more precisely. It also provides a way to easily manage payments and respondent output.[110]

This study recruited 318 respondents, and each was randomly assigned to one of the three conditions. The study only recruited those individuals who indicated that English was their first language, who had written an online review for a product or restaurant in the past six months, and who lived in the United States. The demographics of the sample are presented in Table 1 below.

Table 1: Sample Demographics

Gender	**Percentage of Respondents**
Female	49%
Age Bracket	
18–25	11%
26–35	36%
36–45	27%
46–55	14%
56–65	8%
66+	3%
Education	
2-year degree	11%
4-year degree	43%
Doctorate	3%
High School Grad	8%
Professional Degree	16%
Some College	19%

2. Procedure

After recruitment, respondents were brought to the starting page of the study. On this page, they read the directions of the study and a consent form. Respondents were then asked some gating questions. First, they were asked an attention check question, which ensures that

110. Litman et al., *supra* note 106, at 438, 440.

respondents are not just clicking through the study but are carefully reading directions.[111] Those who failed the attention check were immediately brought to the conclusion of the study, no data on those individuals was collected, and they were not paid for the study.

Second, the respondents answered questions about whether they had written an online review for a product or dining experience in the past six months. They were then asked to provide a description of what platform they used for that review and to provide a link for that review, to ensure that they actually had previously written an online review.

Once the gating questions were completed, respondents went to the recall page. The study employed an aided recall method, where respondents are asked to recall how they experienced and felt during a particular event in their lives. This a common method in consumer psychology research to get respondents to write extensively about a previous event, put respondents in a certain mood, or have them comment on that experience.[112]

This particular study prompted respondents to recall a negative dining experience. Respondents saw the following prompt below on the aided recall page, and they were forced to stay on the page for at least thirty seconds prior to the "next" button appearing to ensure that they read the directions carefully:

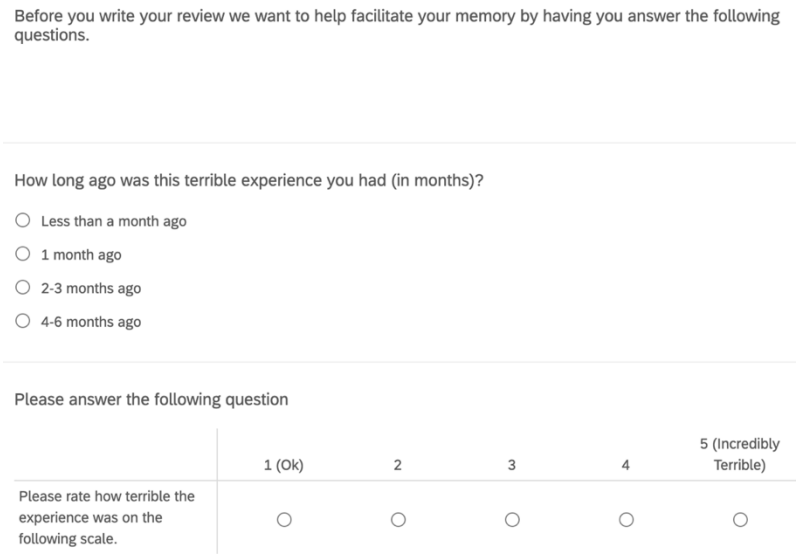
> We would like you to now recall a terrible experience you have had in the past six months with any kind of dining establishment (fast food, fancy restaurant, food hall, food truck, etc.) This could include your experience eating at a restaurant, receiving delivery, or picking up take-out.
>
> Think about the full process of ordering the food to finally consuming it, including how difficult it was to order, the quality of service you received, the price of the meal, the presentation of the meal, and of course the taste of the meal. Focus not just on the events that

111. Attention checks are a very common way for studies to disqualify individuals who are not taking the study seriously. *See generally* James D. Abbey & Margaret G. Meloy, *Attention by Design: Using Attention Checks to Detect Inattentive Respondents and Improve Data Quality*, 53–56 J. OPERATIONS MGMT. 63 (2017); Adam J. Berinsky, Michele F. Margolis & Michael W. Sances, *Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys*, 58 AM. J. POL. SCI. 739 (2014).

112. This often occurs in marketing scholarship where individuals are asked to recall certain advertisements or brand experiences. *See generally* Brian D. Till & Daniel W. Baack, *Recall And Persuasion: Does Creative Advertising Matter?*, 34 J. ADVERT. 47 (2005); George M. Zinkhan, *An Empirical Investigation of Aided Recall in Advertising*, 5 CURRENT ISSUES & RSCH. ADVERT. 137 (1982); May O. Lwin & Maureen Morrin, *Scenting Movie Theatre Commercials: The Impact of Scent and Pictures on Brand Evaluations and Ad Recall*, 11 J. CONSUMER BEHAV. 264 (2012).

> happened, but also how you felt during the process and how you felt after you finished your meal.
>
> Take a few minutes now to vividly imagine this terrible experience again. You will be asked to write about it in the next section.

After spending time recalling the negative experience, respondents were then asked questions about this experience. These questions were intended to further solidify the moment in the respondents' minds so they could easily write about it later in the study. Figure 1 below reproduces these questions.


Before you write your review we want to help facilitate your memory by having you answer the following questions.

How long ago was this terrible experience you had (in months)?

- ○ Less than a month ago
- ○ 1 month ago
- ○ 2-3 months ago
- ○ 4-6 months ago

Please answer the following question

	1 (Ok)	2	3	4	5 (Incredibly Terrible)
Please rate how terrible the experience was on the following scale.	○	○	○	○	○

Figure 1: Aided Recall Questions

After answering these questions, respondents were brought to a page where they were asked to write about the negative dining experience like they were writing an online review. At this point, each respondent had been randomly assigned one of the three censorship conditions (no censorship, specific censorship, or broad censorship).[113] The text box for inputting the review had a minimum character count of seventy-five, which ensured that respondents did write something rather than just click through the negative review part of the study.

113. In actuality, there were five conditions (no censorship, specific censorship with explanation, specific censorship without explanation, broad censorship with explanation, and broad censorship without explanation). *See infra* note 117.

Once this was completed, respondents answered some demographic questions including age, gender, education level, and the device they took the survey on.[114]

C. Results of the Study

After data was collected, the text of each review was run through the LIWC 2015 text analysis software.[115] The software provided analytics on each respondent's negative online review. These analytics were then broken out by condition and then were compared using standard ANOVA[116] statistical analysis to see if there were differences in the text among the three conditions.[117]

While various text characteristics were analyzed, Table 2 below presents the descriptive statistics of a subset of those that showed some significant or marginal significant difference.[118]

---

114. Respondents were informed not to take the study on a mobile device because mobile users have been shown to write shorter reviews and more data was preferable. Those who did take the survey on the mobile device were excluded from the data analysis.

115. For a discussion of the software, see *supra* note 105 and accompanying text.

116. An analysis of variance test ("ANOVA") is used to compare the average response rate of two groups. The means of each response are compared using the variance of each sample to determine whether the two samples have means that are statistically different from each other. A statistically significant result indicates that the means of the two groups are highly likely to be different from each other. In social science methodology, the level of significance that is deemed to be statistically significant is five percent or one percent (which means that there is a five percent/one percent likelihood of seeing a difference in means between two groups, when in reality the means of the two groups are the same). A ten percent significance is deemed "marginally significant." The level of significance of each test below is designated "p." For a discussion of the ANOVA test, see THOMAS J. QUIRK, EXCEL 2007 FOR EDUCATIONAL AND PSYCHOLOGICAL STATISTICS 163–72 (2012).

117. In actuality, there were five conditions into which respondents were randomly sorted. For each censorship condition, there were two separate further conditions. One was an explained condition, and one was a non-explained condition. In the explained condition, the respondents were told that the reason for the censorship (either specific or broad) was that the text analysis software that would be used to read and analyze the data could not effectively analyze those words. For the non-explained condition, no explanation of the censorship was given, just the censorship itself. The thought was that respondents may behave differently if they were to know that the censorship did not come from a content-based motivation, but rather a practical motivation. However, based upon the subsequent analysis, there was no statistical difference between the two sub conditions. That is, providing an explanation did not affect how respondents wrote their reviews. As such, those sub conditions were simply collapsed into the main conditions, leaving the study with three conditions.

118. For a list of all text measures that were analyzed, see *supra* note 105 and accompanying text.

Table 2: Descriptive Statistics

Characteristic of Negative Review	Condition	N	Mean	Std. Deviation
	No Censorship	100	23.87	24.40
Tone	Specific Censorship	106	29.19	26.05
	Broad Censorship	112	34.64	28.85
	No Censorship	100	1.74	1.53
Positive Emotion	Specific Censorship	106	1.83	1.49
	Broad Censorship	112	2.47	2.20
	No Censorship	100	2.52	1.69
Negative Emotion	Specific Censorship	106	2.35	2.59
	Broad Censorship	112	2.44	1.88

The mean scores for the three characteristics by themselves do not indicate much. Each score is on a different, unrelated scale. Therefore, it does not make sense to compare the positive emotion to the negative emotion for a given text string. Another way to put it is that there is no clean way to say that something is more positive than it is negative. Instead, the best comparison is between conditions to see if there are any differences in the positive or negative emotion given the three conditions. It is important to note that positive and negative in this context are not opposite ends of the scale. Rather, the positive and negative emotion scores are simply measuring how many and to what extent positive or negative words were used in the negative review. In addition, "tone" is a type of composite score that combines the two

variables — positive and negative emotion — into one summary variable.[119] The larger the tone score, the more positive the text is.

Comparing the text of reviews of this study to the historical text of social media usage shows generally that respondents in this study were using the study like they used social media writ large and followed the instructions in the study.[120] The average positive emotion and negative emotion via an analysis of random Twitter posts was 5.48 and 2.14, respectively.[121] Comparing these to the no censorship condition is the best check to make sure our respondents were not unique with respect to most social media users. The average negative emotion presented in the reviews written by those in the no censorship condition was 2.5, which is quite close to the prevailing Twitter average of 2.14. We can expect the negative emotion of the reviews to be higher in this study simply because respondents were asked to recall an extremely negative experience. The positive mean score of the no censorship condition was 1.7 while the prevailing Twitter average was 5.48. This confirms that the reviews the respondents wrote were on average less positive than random Twitter posts. This confirms that respondents understood the instructions and wrote relatively negative reviews.

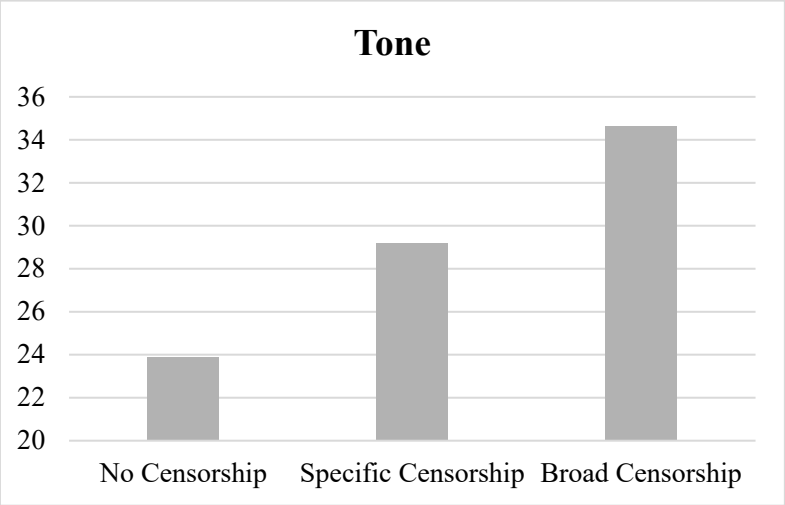
Figures 2, 3, and 4 compare the Tone, Positive Emotion, and Negative Emotion scores for each condition. The figures also include the statistical significance of any differences among the scores.

---

119. *See* Michael A. Cohn, Matthias R. Mehl & James W. Pennebaker, *Linguistic Markers of Psychological Change Surrounding September 11, 2001*, 15 PSYCH. SCI. 687, 689 (2004) (discussing the combining of negative and positive emotion into one variable).

120. LIWC 2015 provides this data to users of their product so that researchers may test whether their studies are replicating sufficiently real-world behavior. *See* PENNEBAKER ET AL., *supra* note 105.

121. *Id.* at 11.



ANOVA Results: No v. Broad (p=0.01); No v. Specific (p=0.032); Specific v. Broad (p=0.28)

Figure 2: Tone Scores by Condition

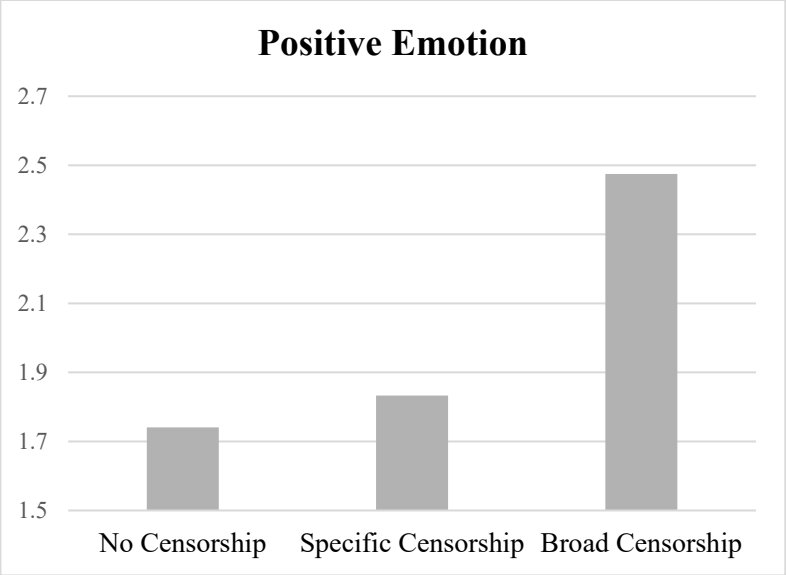
Based upon the above Figure 2, the specific censorship condition created negative reviews that had on average a more positive tone than the no censorship condition (higher score equals more positive tone overall). In addition, the broad censorship condition created negative reviews that had on average an even more positive tone than the specific censorship condition. Looking at the p-values, we see that the positive tone difference between the no censorship condition and the specific as well as the broad condition were significant at the 5% level.[122] Thus, the censorship conditions did chill speech to some degree because the texts of the reviews were different when they were not predicted to be.

Those respondents who were given specific words of censorship and broad categories of censorship wrote reviews that had a more positive tone than those respondents who were given no censorship. Given that both the specific and broad censorship conditions were chosen to not actually limit the language respondents used in the no censorship condition, any change we see is evidence of chilling. Respondents in the face of censorship and potential sanctions felt the need to make their negative reviews significantly more positive — they were chilled in the

122. The general rule of thumb in consumer psychology research is to look for p-values of 0.05 or a significant level of 5%. This means specifically that the chances of seeing a difference like the one above in Figure 2 when really there is no difference at all is 5%. Most scholarship using these methodologies recognizes that a 5% significance level is representative of a meaningful difference.

strict sense of the concept. Although not statistically significant, the broad censorship condition seemed to chill the negative reviews more than the specific censorship condition — which is consistent with the predictions associated with the chilling effect.

The tone metric is made up of a combination of positive and negative emotive words. The figures below analyze the positive and negative emotions present in the reviews.



ANOVA Results: No v. Broad (p=0.007), No v. Specific (p=0.92), Specific v. Broad (p=0.019)

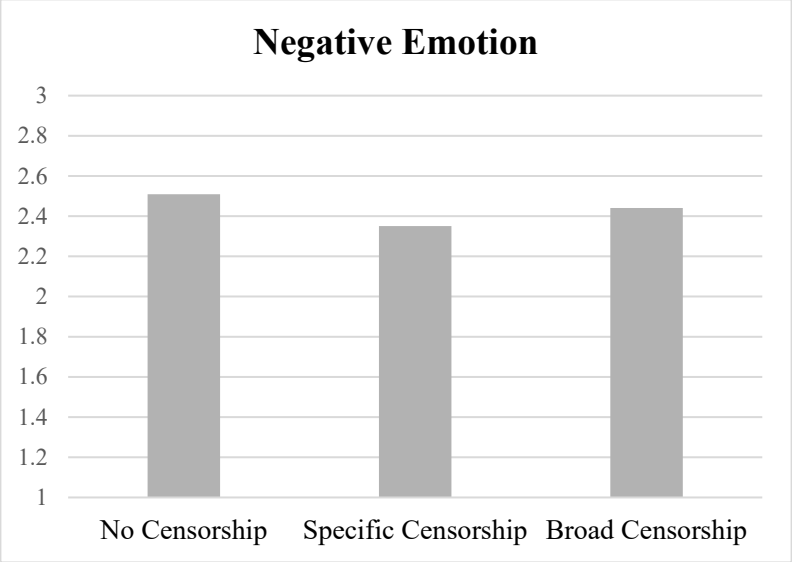
Figure 3: Positive Emotion Scores by Condition

Looking specifically at the positive emotion scores, we see a similar story. Those in the censorship conditions wrote negative reviews that had more positive emotional characteristics than those in the no censorship condition. In particular, the broad censorship condition created much more positive emotion than both the no censorship and specific censorship conditions. In addition, the increase in positive emotion was highly significant at a p-value = 0.007 for the no censorship condition and p-value = 0.019 for the specific censorship condition.

This shows that the broad censorship condition had a much larger chilling effect when looking at positive emotion than the specific censorship condition. This makes sense given that the chilling effect is proposed to occur when a regulation or censorship is broader and more uncertain.[123] In the study, respondents were likely more unsure about

123. *See supra* Section II.A.

how the researcher would apply the standard of "NO Foul language, NO Racially charged language, NO Offensive language." After all, what exactly is offensive language or foul language? The uncertain nature of the censorship was predicted to cause more chilling than more certain (specific) censorship and the results confirmed this prediction. Therefore, when looking at positive emotion, it is quite clear that a broad censorship regime unnecessarily changed speech and hence chilled the negative reviews. Although not specified explicitly here, other characteristics of the reviews were marginally significant.[124]



ANOVA Results: No statistically significant differences

Figure 4: Negative Emotion Scores by Condition

While Figures 2 and 3 tell a story of chilling, Figure 4 does not. Figure 4 shows that the negative emotion present in the negative reviews did not change given any kind of censorship. That is, each set of reviews had just as much negative emotion as the reviews where no censorship was given. This is not what one would expect if the chilling effect was present. Much like positive emotion characteristics, one

124. When comparing the sadness of the reviews, the no censorship reviews were scored as sadder than the broad censorship reviews at a p-value of 0.087. This would indicate that the censorship partly chilled the sadness of the reviews, which is more akin to the type of chilling that scholars and courts have recognized. While the data below shows that negativity did not change, and hence chilling did not occur when looking at the communicative intent of the speech, sadness can be interpreted to be part of the purpose of the speech. If that is the case, then one could plausibly argue that there would be a more traditional chilling effect. However, even though the effect was directional, consistent with chilling, it was not statistically significant.

would expect that the broad censorship condition would show much less negative emotion than the no censorship condition and the specific censorship condition. However, this is not the case. Each condition showed statistically the same negativity. This means that positive tone metric above was only driven by the increase in positive emotion.

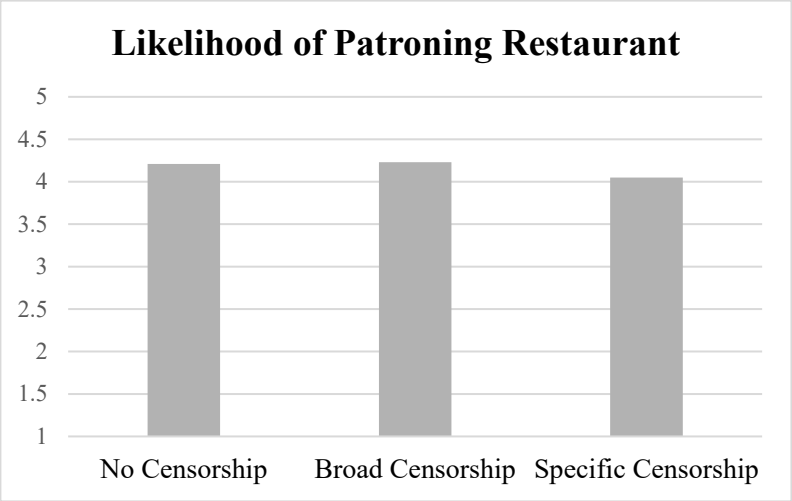
Still, one may observe that if a review has the same level of negative emotion but uses more positively charged words, this may change how the review is perceived.[125] In that case, there would be clear chilling because the message of the reviews would have substantially changed. To determine this, the study randomly selected ten reviews from each of the conditions, a total of thirty reviews. Using a sample recruited from Mechanical Turk,[126] the study asked a separate group of respondents how likely they would be to patronize the dining establishments based *solely* on one review they read.

Respondents (n=146) were randomly assigned to nine reviews (three from each condition) and were asked to rate how likely they would be to patronize the restaurant featured in the review on a 1 to 5 Likert Scale (1 being very likely to patronize and 5 being very unlikely to patronize).[127] Each review received between forty-two and forty-four individual ratings. These ratings were then combined by condition so that each condition had 429 distinct ratings. Figure 5 compares the results of the ratings by condition:

125. I thank Mark Tushnet for raising this concern. The more positive tone as shown in Figure 2 would then lend one to think that the reviews were chilled in a traditional sense. That is, the negative reviews were less negative and hence respondents could not communicate what they wanted to in the sanctioned conditions in comparison to the baseline condition.

126. The study recruited n=146 and only included those respondents who indicated that they sometimes or frequently use online reviews to make purchasing decisions. The sample was 46% female, and the average age range of the sample was 26 to 35.

127. Reviews that included names of the restaurants/dining establishments were redacted so as not to bias any of the respondent's ratings. This ensured that the respondents were just using the text of the review and nothing more.



ANOVA Results: No statistically significant differences

Figure 5: Independent Ratings by Condition

Figure 5 shows that when rating how likely they were to patronize a restaurant based solely on the one review they read, respondents were equally pessimistic of the restaurants in all conditions. That is to say, the reviews in the specific and broad censorship conditions equally persuaded third parties to stay away from the dining establishments. Overall, respondents were highly unlikely to patronize any of the restaurants they read reviews about, which is not surprising given that the reviews were meant to be negative. Therefore, the communicative message of the reviews seemed to stay consistently negative even with the presence of the sanctions.

What does this mean for the chilling effect? The negative reviews seem to have been chilled in the strict sense of the word. Respondents changed their speech (albeit subtly) in the face of censorship and sanctions in ways that the censorship did not intend — they changed their speech unnecessarily by using more positive words and taking a more positive tone overall.

However, the point of the study was to write an extremely negative review. It seems that all the respondents, regardless of the censorship condition, could effectively write the same negatively charged review. The purpose of the speech, to write a negative review, was not chilled by the censorship. The third-party ratings of the reviews confirmed this conclusion. The censorship conditions seemed to have no effect on the respondents communicating how bad their experiences at the dining establishment were.

This study shows that, at least in this context, chilling may exist in the strict sense of the effect, but not in the traditional sense: content moderation designed not to constrain the speech did not show a spill-over chilling effect even though there was some indication that respondents subtly altered their speech. The idea of chilling being problematic occurs because some kind of protected speech that is wanted in the marketplace is deterred. The speech in the study (the negative reviews) did not go away and was not deterred. Instead, the negative reviews in the sanction conditions still equally communicated the negativity that they ultimately intended to portray.

Note that there was not any obvious simple deterrence. When faced with the censorship conditions, respondents did not leave the study. Rather they continued forth even in the face of censorship. This makes sense, as the censorship conditions were chosen so as to not actually affect any of the text of the reviews given in the pilot study. Therefore, when individuals saw the censorship condition, at least consciously, they did not anticipate that the censorship would affect their speech and hence were not deterred in the simple sense.

IV. THE FUTURE OF THE CHILLING EFFECT

The study above is the first to analyze whether and how actual speech gets chilled in the face of regulations and censorship. It showed that speech does get chilled, but not necessarily in the way scholars, courts, and policymakers may expect. Rather than deterring the speech that is intended, the study showed that respondents could still communicate exactly what they want to but instead express it in a more circuitous way. Respondents' negative reviews were just as negative when they faced censorship (hence they communicated exactly what they wanted to); however, they wrote the reviews using more positive emotion (hence they were chilled in at least one way). This Part continues to discuss the results of the study above and puts them in the context of the chilling effect and what it means going forward.

A. The Need for More Empirical Work

At a first pass, there are some weaknesses of the above study and its results that should create the space for more empirical scholarship on the chilling effect. First, as noted above, the censorship took the form of a private speech restriction on social media. Although private speech restriction is important to study for the reasons described above, it should be noted that the chilling effect is mostly used in opposition to government censorship. To the extent that citizen behavior differs in the face of private versus public censorship, this study does not necessarily inform policymakers what the precise effect of a public

regulation will be. This is not to say that private censorship does not track public censorship to some degree, but more work should be done to understand how citizens respond differently (if at all) to public versus private regulations.

Related to this point is the issue around sanctions. Does the extent or severity of the sanction make a difference? According to criminology scholarship, certain criminal sanctions certainly deter more than others.[128] In addition, civil sanctions (e.g., monetary punishment) likely also have different effects than criminal ones (e.g., jail time). Future work should seek to vary the type of sanction to see how that affects the extent of the chilling effect. One would predict that as the severity of sanction increases, measured by both as criminal vs civil and higher versus lower, speech would be chilled even more.

One aspect of this study that is unique is the advent of content analysis. As described above, text analysis was used to analyze the speech of respondents in the study. What this showed is that simply looking at a narrow conception of deterrence (speaking versus not speaking) is not a robust enough way to measure the chilling effect. Speech is complicated and is often contextual, cultural, gendered, racial, etc. Understanding the dynamics of different types of speech and how they change is fundamental to understanding the chilling effect. For example, in the study above, text analysis was used to measure the negativity of each review. In addition, the reviews were evaluated by third parties.

In short, there are many ways to measure how speech changes and, to truly understand how speech gets chilled, scholarship needs to exhaustively seek out these various ways. Only then will we have a more robust conception of whether certain regulations chill speech.

Lastly, this study utilized a social media platform. The context of where speech occurs is likely very important. In this case, respondents likely did not feel that they were unable to express themselves and

---

128. *See generally, e.g.*, Paul H. Robinson & John M. Darley, *The Role of Deterrence in the Formulation of Criminal Law Rules: At Its Worst When Doing Its Best*, 91 GEO. L.J. 949 (2003) (providing a more realistic view of how criminal doctrine manipulation can affect deterrence); Paul H. Robinson & John M. Darley, *Does Criminal Law Deter? A Behavioural Science Investigation*, 24 OXFORD J. LEGAL STUD. 173 (2004) (discussing the extant empirical scholarship on criminal deterrence and finding that deterrence is quite weak at best); Isaac Ehrlich, *The Deterrent Effect of Criminal Law Enforcement*, J. LEGAL STUD. 259 (1972) (modeling empirically the magnitude of deterrence in the face of criminal laws); Raymond Paternoster, *How Much Do We Really Know About Criminal Deterrence?*, 100 J. CRIM. L. & CRIMINOLOGY 765 (2010) (arguing that the criminal justice system is not set up efficiently enough to exploit whatever deterrence the laws create). For legal scholarship focusing on the deterrent effect of civil versus criminal sanctions, see Gary T. Schwartz, *Mixed Theories of Tort Law: Affirming Both Deterrence and Corrective Justice*, 75 TEX. L. REV. 1801 (1997) (arguing that viewing torts as deterrence can bridge the divide between economic and justice-based accounts of torts); W. Jonathan Cardi, Randall D. Penfield & Albert H. Yoon, *Does Tort Law Deter Individuals? A Behavioral Science Study*, 9 J. EMPIRICAL LEGAL STUD. 567 (2012) (showing empirically that tort law does not really deter risky behavior).

likely did not really care if a regulation restricted their speech.[129] In other contexts, like news reporting or tweeting about politics, there may be more emotional reactions to an attempt to restrict speech. These emotional reactions could further manifest the chilling effect or even produce a warming effect where respondents actively ignore and fight back against restrictions. Exploring the various contexts in which speech gets restricted will also likely give more clarity to the chilling effect. In addition, the type of speech activity itself may have implications on the impact of the chilling effect. Speech activity that is focused on reporting facts may be impacted differently in the context of regulation than the speech activity used in the study above.

In summary, this study shows that the chilling effect may exist but, at least in the social media context, may not be as serious as scholars have historically thought. However, this study also shows also how much we still do not know about the chilling effect. Rather, we should be a little more skeptical of both its existence and how it manifests.

B. The Creation of More Civil Speech

It is clear that work on the chilling effect is likely contextual, uncertain, and needs much more rigor. However, this study does give some insights into how social media sites such as Facebook, Twitter, or Instagram should think about potential speech censorship restrictions.[130] Private actors who are contemplating whether and how to limit the amount of speech on their platforms should take pause and think more carefully about the effects of their actions. These actors are faced with arguments that they are silencing protected speech on the one hand and arguments that they are not doing enough to protect individuals from fake news, hate speech, and inappropriate content on the other.

---

129. Of course, they did change their speech, so they did react to it unnecessarily.

130. For discussions on how social media sites are attempting to restrict speech going forward, see Hanna Kozlowska, *Instagram Will Demote "Inappropriate Content" — and Self-Expression Along the Way*, QUARTZ (Apr. 13, 2019), https://qz.com/1594392/instagram-will-demote-inappropriate-content-and-self-expression-along-the-way [https://perma.cc/HPA9-FMCK] (arguing that Instagram's crackdown on inappropriate content will stifle speech); Rod McGuirk, *New Zealand Official Calls Facebook 'Morally Bankrupt'*, ASSOCIATED PRESS (Apr. 8, 2019), https://apnews.com/article/6130548c62654d539c78ba1655a905bd [https://perma.cc/39TU-NKEA] (detailing arguments that Facebook is not doing enough to police speech on their platform); Josh Constine, *Instagram Now Demotes Vaguely 'Inappropriate' Content*, TECHCRUNCH (Apr. 10, 2019, 4:38 PM), https://techcrunch.com/2019/04/10/instagram-borderline [https://perma.cc/FD84-TMSM] (arguing that the uncertainty of Instagram's new policies is likely problematic); Niam Yaraghi, *Regulating Free Speech on Social Media Is Dangerous and Futile*, BROOKINGS (Sept. 21, 2018), https://www.brookings.edu/blog/techtank/2018/09/21/regulating-free-speech-on-social-media-is-dangerous-and-futile [https://perma.cc/NW2N-BUP9] (arguing that these restrictions are inevitably ideologically driven and hence are not beneficial for the marketplace of ideas).

This study can be interpreted to give guidance on censoring speech (in particular hate speech, profanity, or offensive language). It seems that social media users' speech will not necessarily be chilled in the traditional sense, and they will still be able to utilize social media and its functionality for the purposes they want. Rather than necessarily be chilled in the traditional sense, the study implies that social media users can communicate what they want even with restrictions, and when they are restricted, they communicate in a more civil manner. One way to interpret the results is that although the negative contents of the reviews (the purpose of the speech) did not change, how the reviews were written did. They were written in a way that used more positive words and were, therefore, more civil.

So, it stands to argue that when social medial companies limit hate speech and profanity, rather than silencing users, they are creating an environment for users to be more civil and friendly while at the same time still communicating what they want. This is a powerful finding in this context and may give opponents of these social media speech restrictions some pause. If platforms can foster speech with restrictions that is just as good as speech without restrictions and is also less offensive and off-putting, that seems preferable. Certain restrictions may be universally preferred to no restrictions.[131]

C. The Production of More Robust Exchanges

The chilling effect is used to argue against speech restrictions because these restrictions end up limiting the amount of speech in the marketplace of ideas. The results of this study, however, in some ways, turn this argument on its head. Indeed, they may lead one to conclude that chilling actually promotes more speech in the marketplace.

To see this, take the social media context of speech restrictions. Opponents argue that these restrictions harm speech due to the broad content restrictions and sanctions imposed on those that violate these terms. But some scholars have argued that a completely free space to speak is problematic for promoting speech. Hate speech, in particular, has been shown to have serious effects on those individuals at whom it is directed.[132]

The existence of hate speech can cause social harms to "minorities and women by engendering psychological stresses, self-defeating

131. Of course, there are other reasons aside from creating less offensive and more civil speech that social media sites may actually want to restrict speech, including brand building and a focus on a certain stakeholder at the expense of other stakeholders.

132. RODNEY A. SMOLLA, FREE SPEECH IN AN OPEN SOCIETY 152 (1993) (defining hate speech as "the generic term that has come to embrace the use of speech attacks based on race, ethnicity, religion, and sexual orientation or preference").

attitudes, antisocial behaviors."[133] One famous scholar, Richard Delgado, predicts that the promulgation of hate speech causes one of two subsequent behaviors in those minorities that the speech targets: they either respond with hostility or passivity and in turn can add to "children's alienation and sense of rejection."[134] Other scholars have pointed to the "psycho-emotional harms" associated with hate speech.[135] And still others categorize the harm in terms of the transmission model, which focuses on the behavioral and physical responses to hate speech, and the ritual model, which focuses more on the long-term accrued harm of hate speech.[136]

The upshot is that when individuals are presented with continuous instances of hate speech or other harmful language, they often decide to leave productive exchanges. That is, they legitimately change their behavior in ways that are not conducive to reasoned exchanges or simply leave and do not participate in any exchange of ideas.[137] This decrease in the amount of speech actors and hence speech itself is not limited to those that are victims of hate speech. Private companies have threatened to leave social media platforms when those platforms do not restrict offensive and hateful language. In July 2020, over a thousand advertisers threatened to boycott Facebook in response to Facebook's inaction with respect to hate speech.[138] Those advertisers included large accounts like Coca-Cola and Starbucks. Therefore, in addition to individuals leaving productive exchanges due to lack of restrictions, private companies also often leave exchanges, leading to less speech in the marketplace.

The chilling effect that arises in the face of restricting speech can actually stop the exodus of individuals and private companies. Therefore, it can be argued that these restrictions on offensive language, while not affecting the message that individuals communicate, chill speech in a way that makes it more civil and positive. Thereby, restrictions can actually make exchanges on these platforms more robust in terms of increased participation and more civil exchanges.

---

133. Bennett, *supra* note 76, at 474.

134. Richard Delgado, *Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling*, 17 HARV. C.R.-C.L. L. REV. 133, 147 (1982); *see also* N. Douglas Wells, *Whose Community? Whose Rights? — Response to Professor Fiss*, 24 CAP. U. L. REV. 319, 320 (1995) (arguing that hate speech is problematic for society as well as the individuals it is levied against).

135. Toni M. Massaro, *Equality and Freedom of Expression: The Hate Speech Dilemma*, 32 WM. & MARY L. REV. 211, 229 (1991).

136. Clay Calvert, *Hate Speech and Its Harms: A Communication Theory Perspective*, 47 J. COMM. 4, 4 (1997).

137. Ruogu Kang, Laura Dabbish & Katherine Sutton, *Strangers on Your Phone: Why People Use Anonymous Communication Applications*, 19 ACM CONF. ON COMPUT.-SUPPORTED COOP. WORK & SOC. COMPUTING 359 (2016).

138. *Facebook Frustrates Advertisers as Boycott Over Hate Speech Kicks Off*, CNBC (July 1, 2020, 5:56 AM), https://www.cnbc.com/2020/07/01/facebook-frustrates-advertisers-as-boycott-over-hate-speech-kicks-off.html [https://perma.cc/F2UB-PPDS].

If the chilling effect can change the tone of how individuals speak while keeping the content of their speech the same, it could also promote participation in exchanges of ideas from those who would be otherwise offended. The longstanding legal principle then may be reframed as a way to promote more robust speech activity rather than deter it.

V. CONCLUSION

This Article has argued that although the chilling effect is an important and well-utilized argument against both public and private speech censorship, little work has grappled with whether the effect actually exists. This is because measuring the chilling effect is very difficult, and extant work has not utilized all available methodologies.

Using text analysis and experimental manipulation of a social media speech restriction, this Article concludes that, in the context of social media, the chilling effect has little to no impact on the content of the message; at most, it slightly alters the specific style or tone used in speech. The study showed that speech did change in the face of social media censorship — it become more positive and civil. However, the overall communicative effect of the speech did not change — the reviews were just as negative as reviews with no censorship.

This Article then calls into question the realities of the chilling effect and whether the emphasis placed on the effect is really justified. Instead, it argues that much more work needs to be done to better understand how the chilling effect manifests in the marketplace of ideas. Until this is done, policymakers should be skeptical of the force with which opponents of speech restrictions use the chilling effect to invalidate otherwise preferable speech regulations.