Valuing Free Speech: Elite Cues Make Minimal Impact on Partisan American Attitudes about Speech Regulation

Candidate Number: 1078654*

Author: Joshua Berry
MSc in Politics Research
Word Count: 14,077 words

Abstract

Why do Republicans and Democrats in the United States disagree about what type of content, including disinformation and hate speech, should be regulated online? Current explanations for partisan disagreement about disinformation regulation focus on the role of motivated reasoning, internalized preferences (such as moral values), or fact gaps—differences in perceptions about what is untrue—between political partisans. Yet, little, if any, work has been done to understand the impact of in-group elite, partisan cues on preferences for online speech regulation. This experimental study finds that elite cues do not cause consistent heterogeneous effects in shifting polarized partisan policy preferences. Through an experiment showing respondents hypothetical tweets by prominent elites aligned with their own or the opposing party, my findings shed light on the mechanisms driving public attitude formation. Specifically, American attitudes about free speech are durable to one-off elite cues, indicating that partisan divides in preferences for content moderation are likely to result from different sets of ideological preferences weighing the trade-off between individual liberty and collective wellbeing between Democrats and Republicans.

Keywords: Elite Cues; Polarization; Regulation; Online Speech

*Thesis submitted in partial fulfillment of the requirements for the degree of MSc in Politics Research in the Department of Politics and International Relations at the University of Oxford.

1 Introduction

In recent years, increasing media and academic attention in democracies has been devoted to debates about whether it is preferable to preserve individuals' liberty to free speech or protect collective bodies from harm caused by toxic speech. To date, there have been research studies relating to topics of interest within speech: these include discussions about what is fake versus truthful news, when the right to free speech trumps harm-based considerations, the role of disinformation and misinformation as potential threats to democratic elections, and, increasingly, how Internet platforms should police speech or content posted on their platforms (Sunstein, 2022; Guess et al., 2020; Grinberg et al., 2019; Müller and Schwarz, 2021; Greene, 2022; Bazelon, 2022). Political elites have increasingly started discussions about free speech rights, with elites seemingly vacillating between arguing that free speech rights are being infringed upon and that more needs to be done to moderate public speech.

This discussion about free speech, for better or worse, remains firmly attached to American conceptions of free speech. This is largely because most of the web platforms and social media companies that supply the world's global population with Internet chat rooms and algorithmic content feeds are American-domiciled companies. These companies, in turn, tend to be more attentive to the demands of American consumers not only because the American economy is particularly large for multinational corporations but because these staff Americans as their main labor force. Another strong reason as to why American tech companies care about American definitions of free speech is because American-based companies are primarily subject to American law. American law, in turn, is influenced by the voting decisions of the electorate. That is, the attitudes of the American public help determine what the law ultimately determines as permissible versus impermissible speech, which then influences the content moderation policies that American social media companies choose to deploy worldwide. Given the exceptional protection free speech is afforded in the U.S. by American courts, it should be no surprise that there are very little American requirements in terms of what internet content sites can and cannot show (Bollinger and Stone, 2022a). This lax regulatory environ-

ment, fostered by an American law which states internet content companies cannot be held liable for content posted on their website by third-party users, has the American approach to online discourse (Chemerinsky and Chemerinsky, 2022). Some around the world have benefited from this lax regulatory environment, and some have suffered. It is undeniable that there have been unfortunate consequences to individuals brought on by exposure to harmful content online.

These negative consequences have been allowed to spread, partially, because of an online regulatory corpus that has historically been sparse by design. Prompted by the research mentioned above, however, regulators, internet companies, and academics have increasingly focused attention on how best to moderate online content (Gorwa, 2022; Gorwa et al., 2020; Munger, 2017; Grimmelmann, 2015; Kozyreva et al., 2024). New studies, for instance, have uncovered strategies to nudge people into sharing less fake news, analyzed the effect of misinformation on COVID-19 deaths and the 2016 American presidential election, and mapped how hate speech spreads via online networks (Geissler et al., 2023; Guess et al., 2020; Moore et al., 2023; Grinberg et al., 2019; Roozenbeek et al., 2022; Mena, 2020). Studying online content moderation also informs our understanding of new applications of political psychology theory, such as social identity theory, appraisal theory, and moral foundations theory (Sylwester and Purver, 2015; Iyengar et al., 2019; Kosmidis and Theocharis, 2020). All in all, there is a pressing need—and large opportunity—for experimental randomistas to establish why political identity seems to drive polarization in attitudes toward online content moderation.

Today, political partisans possess polarized opinions about appropriate content moderation measures (Pradel et al., 2024; Appel et al., 2023, Gubbala, 2022; Anderson, 2020; Fisher et al., 1999; Balkin, 2022). Research shows that Republicans prefer lower and Democrats higher content moderation (Appel et al., 2023; Pradel et al., 2023). This polarization in policy preferences for regulation presents a challenge as the internet has been used to spread disinformation, hate speech, conspiracy theories (Allcott and Gentzkow, 2017; Saleem et al., 2017; Gallotti et al., 2020). Compared to offline communication, online talk is cheaper, easier, quicker, asynchronous, and more anonymous (Müller and

Schwarz, 2021; Ceresney et al., 2022; Starbird, 2022). Because of these characteristics, researchers theorize that polarizing, fake, and hateful content spreads easier online than offline, driving affective polarization, extremism, and violent political conflict worldwide (SIEGEL and Badaan, 2020; Mathew et al., 2019).

Indeed, a new series of research has focused on examining heterogeneous preferences for internet governance (Pradel et al., 2024; Appel et al., 2023). According to Appel et al. (2023) there are 3 main phenomena that contribute to partisan disagreements over appropriate online moderation:

- Fact gaps ("Differences in perceptions about what is misinformation," including motivated reasoning)
- Party promotion ("A desire to leave misinformation online that promotes one's own party," often called partisan cheerleading)
- Preference gaps (Differences in internalized preferences about whether disinformation should be removed," regardless of whether content favors a political party)

Evidence has been found for all three phenomena (and other factors like emotion, laziness, etc.), but it remains unresolved whether preference gap exists due to (a) differences in internal values, or (b) internalization of elite cues (Appel et al., 2023; Skytte, 2021; Druckman et al., 2013; Pennycook and Rand, 2019; Mosleh and Rand, 2022). In this study, I create an experimental design to adjudicate whether elite cues causally affect partisans' preferences for content moderation.

The question of when and why Americans support free speech rights is a complex one. This study seeks to unravel one specific component of this question by analyzing the profound role of political elite messaging in shaping the political attitudes of the American public. The influence of light elite cues, such as a hypothetical tweet from the Senate leadership of the Republican and Democratic parties, on the political attitudes of members of their respective political parties is a powerful force to be reckoned with.

The findings from this paper are relevant to a pressing public policy issue, contribute to research about theoretical foundations of partisan preference gaps, and inform gaps in

the literature on fundamental differences in how partisans view the regulation of offline versus online speech. Past work, in particular, notes there is insufficient understanding of the impact that elite cues might be playing in affecting individuals stated preferences regarding free speech and policies regarding online content moderation.

My research question is as follows:

Do elite cues causally affect attitudes about content moderation? And, secondarily, does it matter whether moderated information is shared online or in-person?

To answer these research questions, I conduct a randomized online survey experiment where respondents are exposed to hypothetical statements from political elites supporting or opposing two varieties of content moderation, either online or offline moderation. I use six paired experimental groups that expose participants to either a hypothetical Democratic or Republican elite cue about online or offline content moderation. I hypothesize that the medium of shared information matters as I believe individuals intuitively account for fundamental differences between online information (which is cheap, anonymous, and fast) versus information shared in public face-to-face (which is slow and prone to others' counterspeech).

I contribute to filling this gap in empirical work on the role of elite cues in causally affecting Americans' attitudes about free speech by designing an online survey experiment ($N = 3,181$). To answer my research questions above, I conduct a randomized online survey experiment where respondents are exposed to hypothetical statements from political elites supporting or opposing two varieties of content moderation, either online or offline moderation. I use six paired experimental groups that expose participants to either a hypothetical Democratic or Republican elite cue about online or offline content moderation. My experiment exposes a representative sample of American participants gathered via the survey platform Prolific to different versions of hypothetical elite cues from American Democratic Senate Leadership or Republican Senate Leadership that is consistent with theory while taking into account today's real world political climate. Building on research in political science, communication, and behavioral economics, I conceptualize support for free speech as the summative tradeoff between a desire to

protect the societal collective from harm versus preserving the ultimate individual freedom to engage in self-directed action.

My survey experiment reflects current politics by through its use of hypothetical tweets that are based off of political elite statements issued after the first attempted assassination of former American president Donald Trump in the summer of 2024. These political elite statements generally called for American unity and for Americans to be more careful with using inflammatory language. Using these statements as a model, I generate statements that call for American unity and either advocate for greater / less online content moderation / public speech regulation. In other words, I designed my treatments to maximize the likelihood that Republican and Democratic experimental participants would consider their assigned elite cue to be consistent with messaging they had seen from political leaders in the days after former president Trump's shooting. This survey experiment, therefore, also tests the efficacy of unity messaging after the wake of political violence. While this study is focused on the role of elite cues in altering attitudes about speech regulation, its findings can also be extrapolated to other policy debates in which political elites might utilize a unity message to bridge political divides.

The overall story I find from my results is that American views on free speech are rigidly supportive. Regardless of political identification, Democrats, Independents, and Republicans all prioritize free speech over protection from harm, even after exposure to pro-content moderation unity messaging in the wake of newfound political violence. That is to say, I find that exposure to elite cue treatments did not significantly change hearts or minds about supporting greater online content moderation or public speech regulation, whether or not any potential legislation might be considered unconstitutional. This finding aligns with recent studies. When it comes to disagreements about free speech rights or online content moderation, new research suggests there might be more of a perception of partisan disagreement than actual deep divides about embracing free speech as a core American value (Solomon et al., 2024). My study suggests Americans overall tend to be very supportive of free speech rights, regardless of their party identity.

While exposure to elite cues advocating for or against content moderation or public

speech regulation does not lead to any statistically significant effects, I do find that Republicans tend to react to elite cues calling for content moderation by expressing greater disapproval for content moderation and public speech regulation. This differs from Democrats–who typically support content moderation to a greater degree–and Independents–who featured a mixed response of changed attitudes depending on the pro or anti-content moderation content of the message being received.

This lack of consistent significant treatment effects resulting from exposure to elite cues across my dependent variables suggests that American attitudes about free speech are durable to weak and one-off elite cues. While I cannot conclude that attitudes about free speech in the United States are not prone to change through consistent elite campaigning or persuasion attempts, I can conclude that 'breeze in the wind' type cues, such as one-off internet posts, do not consistently change American attitudes. This suggests that American attitudes about free speech might be based on a type of enduring value-based consideration that insulates attitudes from merely following the beliefs of in-group political leaders.

Besides the results involving my main treatments, I find that sociodemographic variables are significant in consistently explaining attitudes about content moderation. Specifically, older Americans, female Americans, and Black Americans are statistically more likely to support content moderation as compared to younger Americans, male Americans, and non-Black Americans. This suggests, particularly in the case of Black and female Americans, that groups at greater risk of being targeted by harmful speech are more supportive of content moderation. Perhaps not surprisingly, my study also confirms via covariate and interaction variable analysis that political ideology and degree of identification with a particular political party can be strong predictors of support for content moderation and public speech laws, although I discover less partisan heterogeneity in my responses than I expected. More important than political identity in explaining attitudes about content moderation is ideology: Americans who generally oppose governmental intervention in societal affairs also significantly oppose speech regulation of any kind.

Lastly, I find that attitudes about the concept of potential public speech laws are harsher than attitudes about online content moderation: Americans almost universally express disapproval for public speech laws of any variety. However, my interaction term analysis shows that attitudes about public speech display greater partisan heterogeneity than attitudes about online content moderation. I find that Republicans, when exposed to any cue involving limitations on public speech, tend to react by further expressing opposition for public speech laws even more fervently than they do for online content moderation.

Overall, the picture this study draws is one where there is limited demand among Americans for greater government involvement in either online or public speech regulation. Indirect, inconsistent, or one-off messaging from political elites in support of content moderation is unlikely to increase American support for legislation that restricts free speech rights.

2 Motivation and Conceptual Framework

2.1 Theory and Hypotheses

While the results of my study have established that one-off elite cues might not play a causal role in altering American's attitudes about speech regulation once established, my original theorizing operated under the theory that elite cues matter in shaping partisans' stated policy preferences. Basing my thinking on theories such as evolutionary psychology and Zaller's theory of the origin of mass opinion, I understood individuals as having some type of political opinion made of a combination of genetic and learned processes (Inglehart, 2018; Zaller, 2006). When individuals are exposed to new information, I theorized they evaluated it by choosing either to accept or reject this new information. If they accepted the new information, they would be forced to update their views to incorporate it. Stronger, more persuasive evidence has a larger and more permanent effect in altering views (Gallotti et al., 2020; Posner, 1998). Stronger evidence can be understood as evidence associated with strong argumentation. That is, strong evidence tends to be coherent, repeated, and said by those who are trusted. Therefore, while elite cues may

provide some information to shift attitudes in a certain direction, elite cues themselves can either be weak types of evidence or strong depending on how it is shared (Nicholson, 2011). When elite cues are repeated over time, they tend to be stronger. When elite cues are mentioned once, as they are seen in this study where hypothetical tweets were shown only once to participants, they are weak. Because I expect elite cues posted online to be naturally weak, I expect elite cues can only dominate prior internal values when (a) my participants' Bayesian prior policy prefernces were very close to a tipping point of supporting versus opposing content moderation policies or (b) when the individual identified so closely with the elite sharing an elite cue that they place much stronger weight on the cue itself.

Based on existing literature, I expect the strength of an elite cue will depend on (1) how forceful the cue is in demanding something and (2) how closely an individual's attitude aligns with that of the person sharing the cue (Nicholson, 2011; Cavaillé and Neundorf, 2023). Those who closely identify with the speaker will alter their priors to a degree greater than that of someone who identifies less strongly with the person who shared a cue. On the other hand, I expect an individual to respond to an out-group cue in much the same manner. Those who weakly identify with an outgroup might ignore the cue or perhaps even slightly modify their priors to accommodate it. Yet those who strongly identify with another political party or group will reject the outgroup elite's cue entirely by doubling down on their previous position or, if they identify so strongly with an opposing political party, by modifying their preferences in the opposite direction of the cue.

Why am I theorizing that ingroup identity matters so much in determining the relative impact of an elite cue? I base this line of logic off of cultural evolution theory which posits that social change evolves in a Darwinian fashion (Henrich, 2015). This theory, derived from the biological sciences, has since been applied to social scientific inquiry, including to study the nature of cooperation and political change (Ostrom, 2015; Inglehart, 2018). Thousands of years ago, our chimpanzee cousins realized the importance of social learning. They found out that in the jungle, the chimpanzees who survived were

those who understood how to use their environment to their advantage, for instance how stones could be used to find food or how branches could be used to sleep safely away from predators. These chimps also saw that the best way to learn these survival tactics was to follow the most successful chimpanzee in their group, often the wisest or biggest of the chimps. By mirroring whatever cues this "elite" chimp did in a day, these younger more naive chimpanzees could figure out how which survival tactics were worth investing in and which were not. And so this trend continued. When early humans banded together in tribes, the value of social learning increased exponentially with our bigger brains and capacity for greater social learning. Still, in these tribes we looked to have our social learning follow the right people. In the interest of our own survival, we decided that the best people to follow were not only the strongest and wisest, but also those that looked like us, thought like us, and acted like us. Those who were similar were a more reliable indicator of what would work for those similar as their life experiences most approximated each other. This trend has been reinforced and reinforced after subsequent periods of human development. Today, where there is more information than ever and decisions about what information to follow comes at a premium, I theorize we lean even more heavily into what cultural evolution has honed for us: following the lead of those elites who are most closely associated with the in-group we most closely associate with (Cavaillé and Neundorf, 2023).

Applying the lens of Darwinian evolution to this process of reinforced social learning provides the core theoretical backing for related types of social theories that involve signalling, such as heuristic/signaling theory (Tversky and Kahneman, 1974; Sniderman et al., 2003) and social identity theory (Tajfel and Turner, 2004). Indeed, Tversky and Kahneman (1974)'s research began a series of studies in behavioral economics. Behavioral economics research established conditions whereby individuals might act in a systematic manner that deviates from behavior consistent with expected utility maximization decisions (Thaler, 2017; Thaler and Sunstein, 2021; Sunstein, 2022). Within the field of behavioral economics, taking heuristics versus using deliberative cognitive processes is a hallmark of quick, intuitive, irrational thought processes–that is a thought process that deviates from what we would normally consider utility-maximizing processes. While

following elite cues are rooted in evolutionary behavior, I argue that allowing an elite cue of the variety shown in this study–that is an explicitly hypothetical tweet–to modify one's political preferences would clearly be an example of irrational behavior. To summarize my theory of how elite cues alter political attitudes, I expect that exposure to one-off elite cues will not influence American attitudes about speech regulation except when affinity to a political party is strong enough to convince an individual to follow the message of their in-group elite, or in cases where American minds are so 'on the fence' about speech moderation as a topic that an elite cue plays a tipping role in convincing individuals to support or oppose a change in speech moderation.

Based on the above theory, I hypothesize (Ha) that political partisans will display a preference to stay consistent with in-group partisan norms to a degree that dominates their prior attitudes about online content moderation. In other words, I hypothesize that in-group elite cues will dominate participants' interval values (contrasting effect), while out-group elite cues will amplify internal values (polarizing effect). In other words. I hypothesize that Preference (P) changes when individuals value $C > V$ where C = Elite Cues and V = Internal Values.

What if my findings are reversed or I find null effects? My null hypothesis ($H0$) is that the public has principles, e.g. that internal values dominate elite cues. Written out mathematically, I can explain this hypothesis as Preference (P) stays stable when individuals value $C \leq V$ where C = Elite Cues and V = Internal Values. If my null hypothesis receives empirical support as I expect, it can be that the public disagreed that elite cues accurately signal in-group preferences. This would be in accordance with mesofoundations theory by Kertzer and Zeitzoff (2017)which holds that when cues aren't seen as elite enough, the public defaults to the dominant laymen position of their in-group.

Yet, what if internal values dominate elite cues? In this case, exposure to elite cues will not alter attitudes in support of or opposing content moderation in a significantly systematic manner. In this case, I expect Americans to be so socialized to prioritize free speech as a core value so as not to be affected by elite cues.

There is a litany of research that suggests American conceptions of free speech do indeed differ from citizens in other democratic countries, including those in Western Europe. For instance, research shows that Americans tend to greatly value free speech, with American progressives and conservatives both believing that free speech is a value consistent with their political ideologies (Mutz, 2023). Further historical research traces the history of the concept of free speech in the United States from its adoption as the First Amendment in the U.S. Constitution by the Framers to free speech's expansion post World War I through concentrated court action that sought to entrench free speech rights that prevented government from abridging a liberal conception of citizens right to free speech (West, 2004; Richards, 1974; Alvarez and Kemmelmeier, 2017). Much of today's concept of free speech rights in the United States is directly the result of court cases decided by the American courts and legal theorizing of American lawyers and judges when deciding these cases.

Indeed, much of the American concept of free speech today can be thought of as generated via Americans' general fascination with precedent and law, as well as Americans' persuasive distrust of government (Wright, 2019; Krotoszynski, 2015). This is why much of the research conducted today has been conducted not by political scientists or economists, but by legal scholars based at law schools who either investigate the philosophical underpinnings of free speech law, or investigate how past court cases have shaped legal precedent in creating cultures of free speech. Comparative legal scholars have found a number of research findings that suggest that Americans contemplate free speech as a uniquely proud cultural value to endorse, whether or not their conception of free speech aligns with the actual definition as established through legal precedent (Wright, 2019; Redish, 1982). Research shows, for instance, that Americans tend to posses a belief that the right to individual liberties justify free speech, while European legal justification for free speech instead tends to emphasize that free speech is a necessary condition of enabling human dignity (Carmi, 2008). Americans, on the other hand, barely consider human dignity as a factor in calling for free speech and indeed human dignity has scantly been cited by the Supreme Court as a rationale for why speech should be protected (Carmi, 2008). Compared to European countries, the United States

has also had conceived as free speech as a powerful limit in constraining the power of a federal government that might engage in tyranny, meaning that Americans may posses a larger conception of threat when their free speech rights appear to be abridged (Krotoszynski, 2015). The different conception of free speech rights in the United States compared to elsewhere in the world, largely a factor of Americans obsession with the Constitution and legal precedent, essentially creates a scenario where the legal right to free speech has also create a cultural value associated with supporting free speech. Some authors, in fact, have claimed that there is no more important concept to American political identity than advocating for free speech (Volokh, 2023). This socialization trend of law influencing American cultural values continues today as prominent law professors have advocated for creating greater cultures of free speech in the classroom, which they hope can inspire future lawyers to defend and promote the cause of free speech in subsequent American free speech cases (Volokh, 2023).

As to why the American legal system has evidently become infatuated with cases relating to free speech, all one has to do is analyze the special role that has been assigned to classified speech in the American legal system. American law is very explicit in saying free speech and free speech legal cases are classified under a different standard of legal scrutiny as they are classified under Strict Scrutiny standards under the First Amendment (Kramer, 2022; Persily, 2022; Klobuchar, 2022; Whitehouse, 2022; Ceresney et al., 2022; Strauss, 2022; Lessig, 2022; Lakier, 2022). In other words, cases dealing with free speech require the courts to analyze the case in a narrow fashion only (Douek, 2022). While two speech cases might appear to be on the same topic, the courts may decide instead that the cases are sufficiently distinct to mandate different court cases and different considerations of the speech involved. This process has, in part, contributed to the relatively recent deluge of free speech cases that the court is analyzing in regard to online speech and internet content as speech. Prominent legal scholars have theorized that the courts are incredibly limited in their ability to issue rulings advocating for state intervention in bringing about greater content moderation online (DiResta, 2022; Bollinger and Stone, 2022b; Adams et al., 2022). Ironically, the current strong protection for free speech that the court has become constrained in making any significant ruling

on free speech is due largely by the courts' own efforts to entrench free speech rights in the 20th century (Franks, 2022; Chemerinsky and Chemerinsky, 2022). Constrained by historical interpretations of free speech law, many prominent legal theorists say that Section 230 of the Telecommunications Act of 1996 (which reads, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.") is so strong in providing immunity for online platforms for the third-party content posted on their websites that there, legally speaking, is very little the government could conceivably do in increasing online content moderation minus the creation of new legislation (Whitehouse, 2022; Congress, 1996). Among the more popular arguments for how to respect Section 230 while creating avenues for greater online content moderation is to allow consumers to sue online platforms as liable for the harm, the consumers received through harmful viewing content on their platforms (Adams et al., 2022). Needless to say, though, legal scholars concur that the strong legal precedent of free speech in the United States and the general American populace's (1) awareness that free speech is protected strongly by the American courts and (2) adoption of the court's language of historical precedent in analyzing any changes to constitutional rights, has created an environment where Americans from judges to laypeople strongly value free speech as a cultural value. That is to say, legal theorists maintain that Americans internally value free speech as a core American value to such a degree that judges, who are among the most educated citizens in the country and are responsible for dispensing justice, generally work to prevent activities that limit free speech rights as they understand it.

So, if American attitudes about speech regulation do not change after exposure to elite cues, I can conclude that (1) either my elite cues failed to elicit a result because of an error in my design or (2) that Americans value free speech to such a degree as to ignore elite cues that advocate for reduced speech rights. This section's discussion of when I theorize elite cues will shift attitude, as in cases where people identify strongly with a certain political party, versus why I believe elite cues may not shift attitudes (due to the strong cultural esteem held by Americans for free speech as a positive liberty-granting value) provides rationale for my null and alternative hypotheses discussed in

this section.

To recap, based on analysis of my motivating theory, I derive the following hypotheses:

$H0$: internal values dominate elite cues. I expect Preference (P) to stay stable when individuals value $C \leq V$ where C = Elite Cues and V = Internal Values. This result can come about either due to elite cues (C) being weak due to lack of credibleness, minimal persuasive evidence provided through a one-off treatment, or internal values (V) being strong enough to resist exposure to elite cues. These internal cues might be strong due to Americans embracing freedom of speech as a cornerstone cultural value. If this hypothesis is true, I do not expect to find changes in attitudes within partisan groups, as revealed through analysis of partisan interaction effects.

Ha: elite cues dominate internal values. I expect Preference (P) for online content moderation or public speech regulation changes when individuals value $C > V$ where C = Elite Cues and V = Internal Values. I most expect this result to occur among Americans who feel the greatest affinity to a political in-group (such as Americans who self-identify as Strong Democrats or Strong Republicans) as these Americans are likelier to follow a heuristic of accepting elite cues. I also expect this alternative hypothesis to be proven correct if it turns out that Americans have weak internal values driving weak preferences for free speech rights. My hypothesis test is to analyze changes in attitudes within partisan groups, as revealed through analysis of partisan interaction effects.

Because I suspect that internal values regarding face-to-face public speaking will be stronger than those regarding online content moderation given the long legacy of American court precedent protecting free speech in face-to-face settings, I expect that ($H1$) changes in preferences for public speech regulation within political identity groups will be affected to a lesser, or perhaps even negative degree, as compared to changes in preferences for online content moderation. As with all my hypotheses, I am interested in analyzing changes in attitudes within political identity groups.

3 Research Design and Methodology

To measure the causal effects of elite cues calling for either greater or less online content moderation and public speech regulation in the United States, I randomly exposed people to hypothetical social media posts from Democratic Senate Leadership or Republican Senate Leadership. I opted for a design where I operationalized my six treatment groups to either expose participants–who were asked to identify themselves as political Democrats, Republicans, or Independents–to a hypothetical tweet from (1) Republican Senate Leadership or Democratic Senate Leadership and (2) that either was pro-online content moderation, anti-online content moderation, or pro-offline content moderation. These tweets were modeled after real statements, press releases, and tweets released by American political leaders after the assassination attempt against former President Donald Trump on 13 July 2024 (Scherer, 2024; Scalise, 2024; Biden, 2024a; Trump, 2024; Johnson, 2024). These real tweets echoed statements put out by former President Trump and sitting President Joe Biden by calling for American "unity" and a reduction in "incendiary language" (Biden, 2024b; FitzGerald, 2024). Analysis from my pilot study indicated that participants thought my tweets were realistic. My treatment conditions are designed to evaluate the importance of (1) in-group, partisan elite signals in mediating prior attitudes about value-based policy positions and (2) whether the context of information (i.e., online versus offline information) matters to the public when evaluating the validity of free speech as a right. Participants randomized into a control group were not exposed to any tweets.

My overall design conditions can be described as follows. While I originally intended to run a blocked version of my experiment where I excluded political Independents from my analysis or only exposed participants to their in-group elite cues, I exposed Democrats, Republicans, and Independents to both their in-group and out-group elite cues. Therefore, I randomly assigned participants to one of six treatment groups or one control group.

I recruited a representative sample of American participants using the Prolific platform. To ensure my design operated as intended and that participants were able to participate

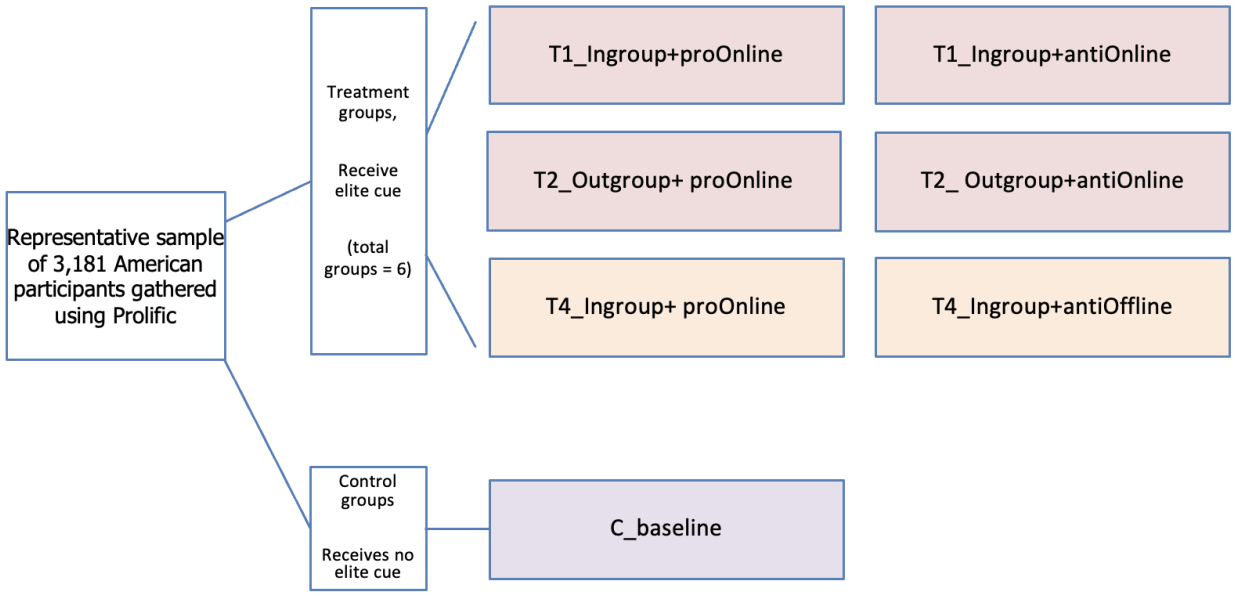


Figure 1: Experimental Design

in the study without any barriers successfully, I first launched a pilot study for 300 representative American participants. Finding no issues with my design, I rolled my pilot study into my full experiment, where I deployed the remainder of my research funding to recruit an additional 2,881 participants. Because my pilot launch only differed from my full launch by a final qualitative question asking for participant feedback as the last question after they had completed the experiment and agreed to submit their responses, I was able to use my pilot data in my analysis, bringing up my number of participants to 3181 representative Americans.

Unfortunately, due to funding constraints, this number of participants is less than the 3300 participants I calculated as required through my power calculations. My representative and balanced sample likely aided, though, in increasing my survey power. Please refer to the Appendix for further information on power calculations. Because I am 119 participants short of the amount I calculated as required to identify effects with 0.8 power at the 95% confidence level, the findings shared in this thesis should be considered like a pilot. I will seek more funding to complete participant recruitment and make further improvements to this design possible in the coming year.

3.1 Treatments

Participants first completed a Qualtrics-designed pre-treatment survey that gathered demographic variables as well as a pre-treatment measure of my main dependent variables, a feeling thermometer of support for online content moderation and a feeling thermometer for public speech regulation. While I was aware this pre-treatment measure might create a participant anchoring effect, I was careful to vary the language of my pre and post-treatment feeling thermometers. Moreover, I was interested in collecting this pre-treatment measure so I could analyze how exposure to elite cues shifted pre-existing attitudes through a pre-post model that subtracted pre-treatment support for speech regulation from post-treatment support. This measure would reveal whether elite cues alone were sufficient in changing internal values primed by the pre-treatment measure. My pre-treatment survey also collected political identity on a seven-point scale, allowing me to analyze how identity proximity to party elites might alter the effect of elite cues on participant attitudes.

After this pre-treatment survey, control group participants proceeded directly from the pre-treatment questionnaire to the post-treatment questionnaire after a brief clickthrough screen. Treatment group participants were sorted into one of six treatment groups: $T1\_Dem + pro\_Online$, $T1\_Dem + anti\_Online$, $T2\_GOP + pro\_Online$, $T2\_GOP + anti\_Online$, $T4\_Dem + pro\_Offline$, or $T4\_GOP + pro\_Offline$.

These groups, and the acronyms assigned with them, can be understood to mean the number of the treatment group (e.g., T1), whether the treatment group is assigned to Republican or Democrat elite (e.g., Dem), whether the treatment is pro or anti content moderation (e.g., pro), and whether the treatment group asks about offline or online content moderation (e.g., online). See the images below for my treatments as they appeared to participants. Note that I explicitly told my participants that these tweets were hypothetical and were generated based on real statements. I decided not to deceive my participants because past research has established that scenario-based experiments that measure participants' *a priori* intentions can accurately mirror real-world behavior (Weyrich et al., 2020; Rungtusanatham et al., 2011).

Figure 2: Treatment for Democrat Pro Online Content Moderation

Figure 3: Treatment for Democrat Anti Online Content Moderation

Figure 4: Treatment for Republican Pro Online Content Moderation

Figure 5: Treatment for Republican Anti Online Content Moderation

Figure 6: Treatment for Democrat Pro Offline Content Moderation

Figure 7: Treatment for Republican Pro Offline Content Moderation

3.2 Post-Treatment Survey

I tested my primary hypotheses by capturing a series of outcome variables, including primary feeling thermometers for online content moderation and public speech regulation. My post-treatment survey also captured other outcomes related to legal preferences, judgments on the fairness of past Supreme Court cases related to speech, and which values participants considered when determining policy preferences for speech regulation. My questions were generated from my overarching theory and can be seen in full in this thesis' appendix. It is worth noting that the post-treatment feeling thermometers differed in wording from my pre-treatment measures and were spaced out in my survey so as to minimize anchoring effects from capturing my main outcomes both prior to and post treatment.

In summary, my key research variables are as follows. My independent variable (IV) was elite cues for and against either online content moderation or public speech regulation through law. My pre-treatment survey gathered demographic variables that I added as covariates in my analysis. My dependent variable (DV) was partisan preferences for and attitudes about content moderation. While a full list of survey questions I asked can be found in my appendix, I generally designed my post-treatment survey to test for my theory-guided hypotheses, seeking to ascertain the trade-off participants saw between free speech rights and collective protection from harm driven by varieties of speech. I interacted participants' political identity and also their strength of political identity with treatment groups for subgroup results to calculate conditional average treatment effects (CATE).

4 Results

My analysis studies whether the identity of an elite cue provider and the area of speech regulation (online versus offline moderation) causally alter partisans' Bayesian priors about the permissiveness of speech regulation. Following the experimental analysis convention, I model a continuous outcome for my outcomes as a function of six treatment

conditions compared to a control group, and I interact treatments with participants' political party affiliations. The most simplified specification for my main feeling thermometer outcomes is modeled as regressing the subject's support for speech regulation (outcome) on individual treatment status interacted with partisanship. I am interested in analyzing the effect of elite cues within partisan groupings, as shown via interaction effects.

$$P(Outcome)i = \beta 0 + \beta 2aT2ai + \beta 2bT2bi + \beta 2cT2ci + \beta 2dT2di + \beta 2eT2ei + \beta 2fT2fi +$$

$$\beta 3aT2aiPi + \beta 3bT2biPi + \beta 3cT2ciPi + \beta 3dT2diPi +$$

$$\beta 3eT2eiPi + \beta 3fT2fiPi + \omega Xi + \epsilon i$$

where:

- i index individuals
- $Outcomei$ has a value of 0 if the individual is fully against the permissiveness of regulators to moderate content and has a value of 100 if the individual is fully in support of the permissiveness of regulators to moderate content
- $\beta 0$ is the probability of supporting online content moderation following exposure to the control group.
- $\beta 2a$ is the treatment effects of being exposed to an IN-group elite cue advocating FOR content moderation in the arena of ONLINE content.
- $\beta 2b$ is the treatment effects of being exposed to an IN-group elite cue advocating AGAINST content moderation of ONLINE content.
- $\beta 2c$ is the treatment effects of being exposed to an IN-group elite cue advocating

FOR content moderation in the arena of OFFLINE (in-person) content.

- $\beta 2d$ is the treatment effects of being exposed to an IN-group elite cue advocating AGAINST content moderation of OFFLINE (in-person) content.
- $\beta 2e$ is the treatment effects of being exposed to an OUT-group elite cue advocating FOR content moderation of ONLINE content.
- $\beta 2f$ is the treatment effects of being exposed to an OUT-group elite cue advocating AGAINST content moderation of ONLINE content.
- $\beta 3aPi$ is the treatment effect of $\beta 2a$ interacted with Political Identity at the individual level, where political identity is measured as the degree to which one affiliates with their political party or a grouped trichotomy of Democrat, Republican, or Independent.
- $\beta 3bPi$ is the treatment effect of $\beta 2b$ interacted with Political Identity at the individual level.
- $\beta 3cPi$ is the treatment effect of $\beta 2c$ interacted with Political Identity at the individual level.
- $\beta 3dPi$ is the treatment effect of $\beta 2d$ interacted with Political Identity at the individual level.
- $\beta 3ePi$ is the treatment effect of $\beta 2e$ interacted with Political Identity at the individual level.
- $\beta 3fPi$ is the treatment effect of $\beta 2f$ interacted with Political Identity at the individual level.
- ωXi are covariate controls such as age, ethnicity, generalized political ideology regarding government intervention, gender, affinity (or self-identified closeness) to an American political park, and education at the individual level.
- ϵi is the error term with zero mean.

I modeled my outcomes using linear regression models, but I include statistical analysis

using logistic regression and multinominal logistic regression modeling in the Appendix for relevant ordered factor and nominal outcome variables. There were no systematic differences when comparing results from my simple linear regression to results generated with other types of regression modeling.

Most broadly, I find that American views on free speech are rigidly supportive. Democrats, Independents, and Republicans all prioritize free speech over protection from harm, even after exposure to pro-content moderation unity messaging in the wake of newfound political violence. Indeed, analysis of my main feeling thermometer measures shows that no treatment group x partisan identity interaction resulted in participants consistently changing their attitudes about speech regulation. That is, within partisan groupings, I found no significant and systematic changes in attitudes across my outcome variables.

The lack of effect for these treatments held across four different model types with increasingly sophisticated specifications (simple regression, regression with covariates, simple interaction regression, interaction regression with covariates). Moreover, I did not find any effect of treatments interacted with political identity for either preferences about online content moderation or public speech regulation.

When analyzing my pre-post models for the effect of my treatment on attitudes, I likewise find no consistent patterns of significant effects in either preference for online content moderation or public speech regulation when compared to the control group.

Because treatments are not significant, I can rule out *Ha* as a hypothesis, and instead accept my null *H0*, that elites cues do not overpower preexisting attitudes, or the preexisting policy preferences that Americans possess about content moderation. This could either be because my cues were too weak to elicit attitudinal shifts, or preexisting attitudes are so strongly held as to resist influence from elite cues. While my results are insignificant, it is worth noting that Republicans tended to react to elite cues calling for content moderation by expressing greater disapproval for content moderation and public speech regulation across cases. This differs from Democrats, who typically supported content moderation to a greater degree post-treatment, and Independents, who reacted to treatments with more variance depending on the pro or anti-content moderation con-

Comparison of Regression Models of Elite Cue Treatment Group Against Favorability Towards Online Content Moderation

Feeling Thermometer of 0 (Coldest) to 100 (Warmest)

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	79.81***	67.53***	77.66***	11.88*
treat				
Control	—	—	—	—
Dem_Offline_Pro	-1.283	-2.437	1.309	4.301
Dem_Online_Anti	-1.795	-1.249	1.042	-1.099
Dem_Online_Pro	-1.735	-2.344	0.7416	2.731
GOP_Offline_Pro	-1.232	-3.563	0.2684	0.8450
GOP_Online_Anti	-0.9218	-1.451	2.117	4.806
GOP_Online_Pro	-1.166	-0.4122	1.417	3.053
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-16.41***	-4.242	-15.13***	-2.433
Republican	-24.10***	-6.562	-18.82***	-1.923
Party_Affinity		1.631		-1.647*
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		-20.95***		-1.792
Gender				
Female		—		—
Male		-6.613***		0.7784
Non-binary		-5.733		3.098
Ethnicity				
White		—		—
Asian		-0.9592		-0.2833
Black		6.322**		2.402
Latino		2.235		3.630
Multiracial		-4.742*		0.9279
Native American		-2.246		1.943
Age		1.735***		0.5382
Socials				
At least once a week but not every day		—		—
A few times a month		0.0269		1.964
Every day		-0.2115		-1.903
Less often or not at all		1.349		-4.180
Community				
Suburb		—		—
City		2.555		0.9727
Other		-3.680		-4.924
Rural		-1.440		0.0188
Town		1.636		2.761
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		-0.0634		0.1795
Graduate degree		-0.8051		-0.2339
Prefer not to say		25.77		-1.254
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			-2.002	-4.774
Dem_Online_Anti * Independent			-0.2533	2.831
Dem_Online_Pro * Independent			1.439	-3.387
GOP_Offline_Pro * Independent			-0.2179	0.2344
GOP_Online_Anti * Independent			-0.9085	-6.562
GOP_Online_Pro * Independent			-6.692	-2.889
Dem_Offline_Pro * Republican			-6.007	-8.296*
Dem_Online_Anti * Republican			-8.049	-1.906
Dem_Online_Pro * Republican			-8.745*	-6.534
GOP_Offline_Pro * Republican			-4.208	-7.868
GOP_Online_Anti * Republican			-8.151	-10.08*
GOP_Online_Pro * Republican			-1.705	-3.491

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 8: Comparison of Post-Treatment Online Content Moderation Preferences

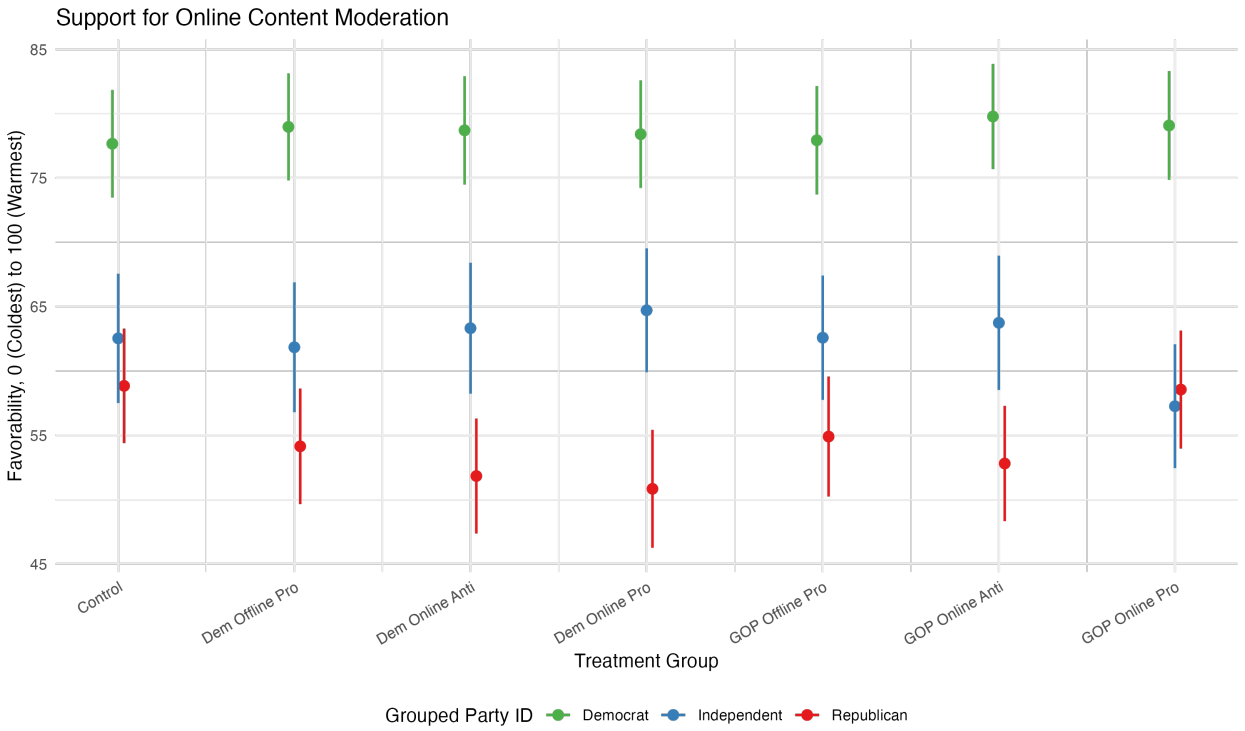


Figure 9: Post-Treatment Online Content Moderation

tent of the message being shared.

Besides my treatments, I find that sociodemographic variables are consistently significant in explaining attitudes about content moderation. Specifically, it appears as if older, female, and Black Americans are statistically more likely to support content moderation. This suggests, particularly in the case of Black and female Americans, that demographic groups that researchers believe are generally at more risk of being targeted by harmful speech, such as toxic speech and hate speech, are more supportive of free speech abatement in favor of protection from harmful speech.

Further debunking my alternative hypothesis, I find that political identity and, even more notably, affinity to one's political party are insignificant in explaining any changes in support for content moderation or public speech laws.

The lack of consistent significance of a 7-point scale that measures party identity affinity as an interaction term with my treatment challenges a core part of my original theory. I had expected that identity-based closeness to a political party would mean greater

Comparison of Regression Models of Elite Cue Treatment Group Against Favorability Towards Public Speech Regulation

Feeling Thermometer of 0 (Coldest) to 100 (Warmest)

	No Covariates			With Covariates			Interaction No Covariates			Interaction With Covariate		
Variables	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**
(Intercept)	66.30	62.96, 69.63	<0.001	53.86	40.33, 67.38	<0.001	64.85	60.03, 69.68	<0.001	52.34	38.12, 66.56	<0.001
treat												
Control	—	—		—	—		—	—		—	—	
Dem_Offline_Pro	1.658	-2.604, 5.919	0.4	1.681	-3.080, 6.443	0.5	2.146	-4.658, 8.950	0.5	2.384	-5.340, 10.11	0.5
Dem_Online_Anti	-0.8307	-5.102, 3.440	0.7	-0.3941	-5.224, 4.436	0.9	0.2995	-6.542, 7.141	>0.9	2.874	-5.057, 10.81	0.5
Dem_Online_Pro	-0.7479	-5.001, 3.505	0.7	0.1280	-4.636, 4.892	>0.9	1.910	-4.913, 8.733	0.6	1.971	-5.754, 9.696	0.6
GOP_Offline_Pro	1.649	-2.625, 5.923	0.4	-0.0951	-4.849, 4.659	>0.9	0.1004	-6.751, 6.952	>0.9	-0.4463	-8.214, 7.321	>0.9
GOP_Online_Anti	0.9887	-3.275, 5.253	0.6	0.5791	-4.223, 5.381	0.8	6.630	-0.1189, 13.38	0.054	7.470	-0.2723, 15.21	0.059
GOP_Online_Pro	0.5508	-3.711, 4.813	0.8	2.104	-2.686, 6.893	0.4	2.054	-4.808, 8.916	0.6	1.238	-6.555, 9.030	0.8
Grouped_Party_ID												
Democrat	—	—		—	—		—	—		—	—	
Independent	-18.16	-21.00, -15.32	<0.001	-2.761	-7.814, 2.291	0.3	-19.05	-26.60, -11.50	<0.001	-4.905	-14.41, 4.600	0.3
Republican	-24.80	-27.49, -22.11	<0.001	-0.4267	-8.550, 7.697	>0.9	-19.89	-26.94, -12.85	<0.001	5.017	-5.884, 15.92	0.4
Party_Affinity				2.037	0.1426, 3.932	0.035				1.948	0.0502, 3.846	0.044
Gov												
There are more things that government should be doing				—	—					—	—	
Government is doing too many things better left to businesses and individuals				-29.56	-32.52, -26.60	<0.001				-29.54	-32.50, -26.58	<0.001
Gender												
Female				—	—					—	—	
Male				-8.971	-11.61, -6.335	<0.001				-8.961	-11.60, -6.322	<0.001
Non-binary				-5.437	-17.63, 6.758	0.4				-5.193	-17.41, 7.028	0.4
Ethnicity												
White				—	—					—	—	
Asian				4.024	-2.304, 10.35	0.2				4.278	-2.053, 10.61	0.2
Black				10.54	6.314, 14.76	<0.001				10.59	6.362, 14.82	<0.001
Latino				1.761	-5.483, 9.005	0.6				2.078	-5.181, 9.337	0.6
Multiracial				-2.768	-7.839, 2.302	0.3				-2.437	-7.516, 2.642	0.3
Native American				-3.522	-20.06, 13.01	0.7				-3.717	-20.28, 12.85	0.7
Age				1.400	0.5285, 2.271	0.002				1.435	0.5626, 2.307	0.001
Socials												
At least once a week but not every day				—	—					—	—	
A few times a month				5.079	-2.821, 12.98	0.2				5.056	-2.860, 12.97	0.2
Every day				1.341	-2.462, 5.143	0.5				1.692	-2.117, 5.500	0.4
Less often or not at all				3.765	-2.903, 10.43	0.3				4.237	-2.442, 10.92	0.2
Community												
Suburb				—	—					—	—	
City				-1.399	-4.628, 1.829	0.4				-1.495	-4.727, 1.736	0.4
Other				2.980	-39.48, 45.44	0.9				5.544	-36.94, 48.03	0.8
Rural				-3.415	-6.998, 0.1668	0.062				-3.510	-7.103, 0.0824	0.055
Town				-1.126	-5.430, 3.178	0.6				-1.620	-5.935, 2.696	0.5
Grouped_Education												
High school or less				—	—					—	—	
Bachelor's or Associates degree				-1.237	-4.568, 2.094	0.5				-1.221	-4.561, 2.119	0.5
Graduate degree				1.208	-2.014, 4.431	0.5				1.143	-2.083, 4.370	0.5
Prefer not to say				23.27	-18.76, 65.29	0.3				23.93	-18.15, 66.00	0.3
treat * Grouped_Party_ID												
Dem_Offline_Pro * Independent							3.098	-7.575, 13.77	0.6	1.999	-10.08, 14.08	0.7
Dem_Online_Anti * Independent							2.896	-7.829, 13.62	0.6	2.092	-10.12, 14.31	0.7
Dem_Online_Pro * Independent							1.367	-9.175, 11.91	0.8	4.864	-7.040, 16.77	0.4
GOP_Offline_Pro * Independent							5.892	-4.681, 16.46	0.3	5.571	-6.304, 17.44	0.4
GOP_Online_Anti * Independent							-3.803	-14.55, 6.947	0.5	-7.303	-19.53, 4.930	0.2
GOP_Online_Pro * Independent							-2.956	-13.52, 7.612	0.6	5.391	-6.646, 17.43	0.4
Dem_Offline_Pro * Republican							-3.829	-13.81, 6.150	0.5	-3.186	-14.25, 7.875	0.6
Dem_Online_Anti * Republican							-5.543	-15.53, 4.444	0.3	-10.48	-21.74, 0.7787	0.068
Dem_Online_Pro * Republican							-9.115	-19.16, 0.9311	0.075	-9.063	-20.27, 2.146	0.11
GOP_Offline_Pro * Republican							-0.1251	-10.24, 9.990	>0.9	-3.092	-14.29, 8.108	0.6
GOP_Online_Anti * Republican							-13.78	-23.71, -3.846	0.007	-13.93	-25.03, -2.823	0.014
GOP_Online_Pro * Republican							-1.610	-11.68, 8.462	0.8	-1.247	-12.46, 9.966	0.8

[1] CI = Confidence Interval

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 10: Comparison of Post-Treatment Public Speech Regulation Preferences

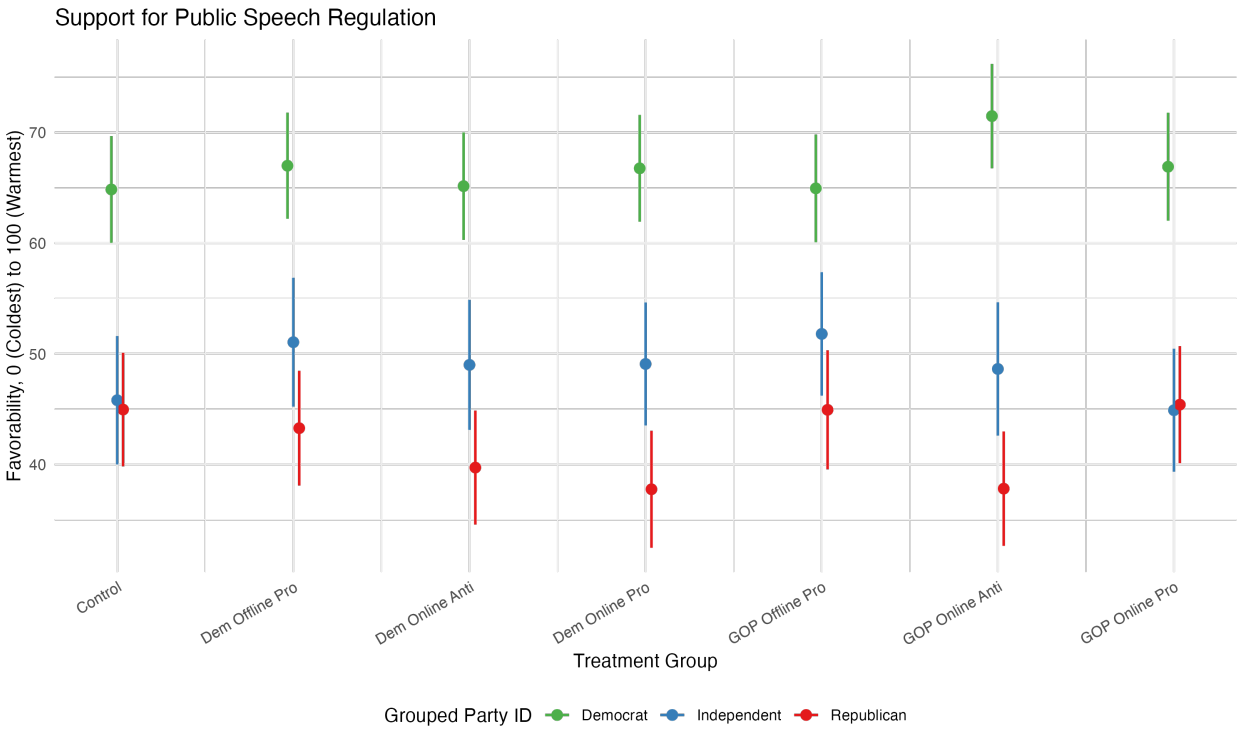


Figure 11: Post-Treatment Public Speech Regulation Preferences

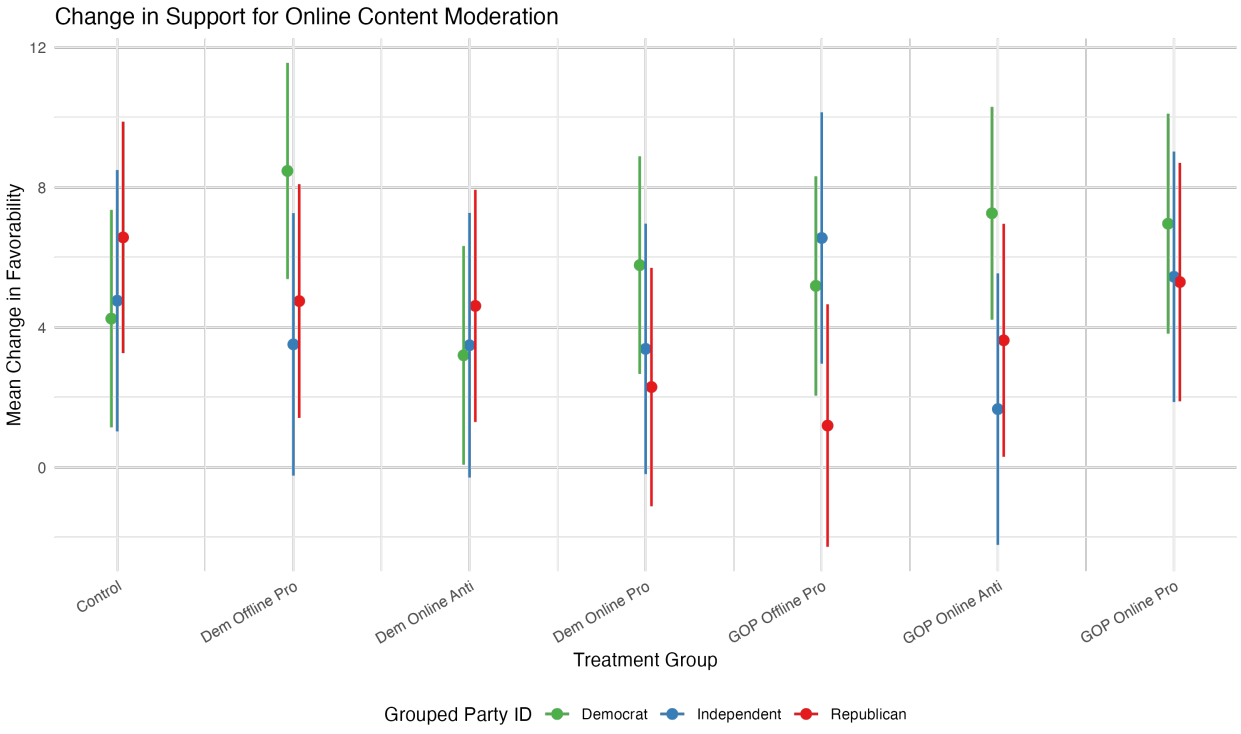


Figure 12: Post-Treatment Pre-Post of Online Content Moderation Preferences

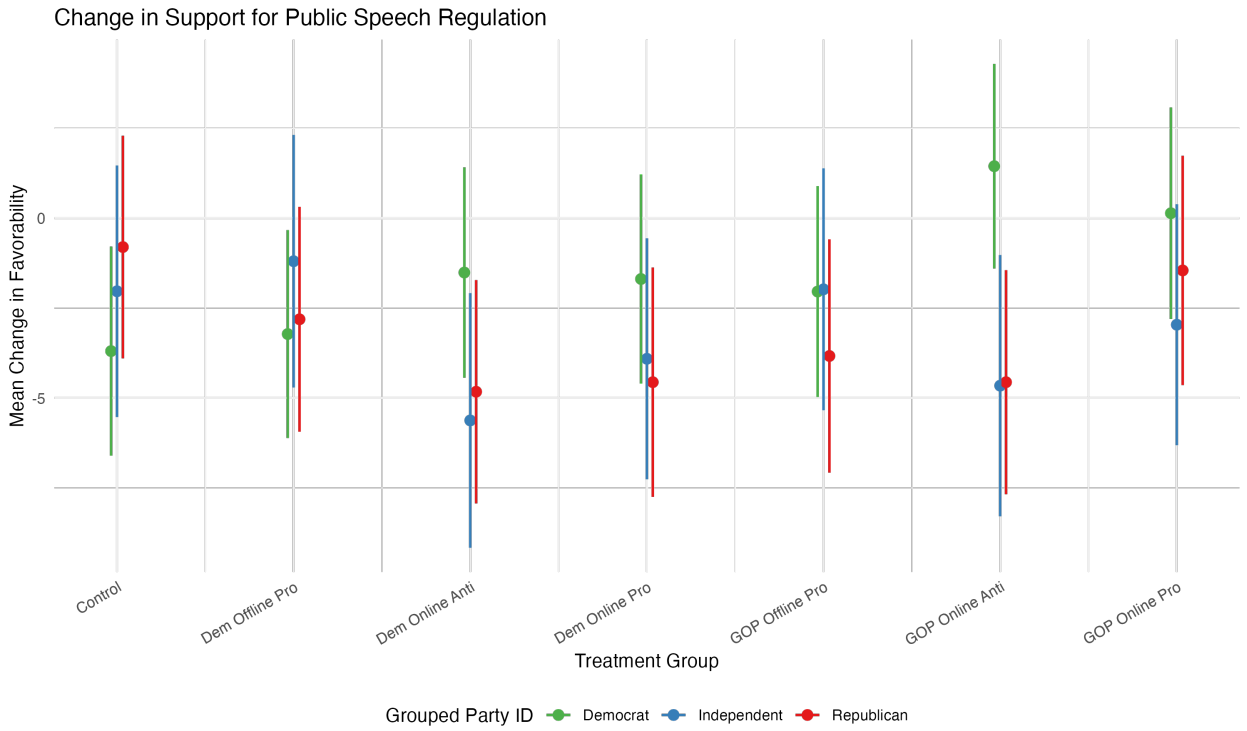


Figure 13: Post-Treatment Pre-Post of Online Content Moderation Preferences

acceptance of elite cues provided by respected, in-group political elites.

Analysis on my main feeling thermometer outcome variables, including my pre-post model outcomes, suggested that Americans' attitudes about laws involving public speech are similar to those about online content moderation in that Americans generally express disapproval for any free speech-abating law, no matter the targeted type of speech. However, I noticed that attitudes about public speech displayed two distinct trends. First, while Americans universally support greater online content moderation when asked about it in the post-treatment questionnaire, no matter their political affiliation, Americans almost always demanded less public speech regulation after being asked about it a second time in my study, no matter their political identity. This suggests that Americans think about public speech regulation in a substantively different manner than online content moderation. Moreover, it suggests that Americans see online content moderation and public speech regulation as disconnected areas of regulation: changing one's mind to support more online content moderation does not entail that Americans will stepwise support greater offline speech regulation. Rather, Americans appear more likely to

Pairwise Contrasts for Interaction Effects regarding Online Content Moderation[1]

Estimated Interaction Effects, Confidence Intervals, and P-values

contrast	treat	Estimated Effect	95% CI Lower	95% CI Upper	P-value
Dem_Offline_Pro - Control	Control	−1.1989	−13.9077	11.5098	0.6093
Dem_Offline_Pro - Control	Dem_Offline_Pro	−1.0791	−13.8793	11.7211	0.5874
Dem_Offline_Pro - Control	Dem_Online_Anti	−0.8788	−13.5734	11.8159	0.5936
Dem_Offline_Pro - Control	Dem_Online_Pro	−0.6407	−13.1007	11.8193	0.6191
Dem_Offline_Pro - Control	GOP_Offline_Pro	−1.0084	−13.7275	11.7106	0.5969
Dem_Offline_Pro - Control	GOP_Online_Anti	−1.3690	−14.2475	11.5095	0.5907
Dem_Offline_Pro - Control	GOP_Online_Pro	−1.5486	−14.4669	11.3698	0.5943
Dem_Online_Anti - Control	Control	−1.3560	−14.2162	11.5042	0.2909
Dem_Online_Anti - Control	Dem_Offline_Pro	−0.3860	−13.3426	12.5707	0.3065
Dem_Online_Anti - Control	Dem_Online_Anti	−0.3684	−13.2125	12.4756	0.2970
Dem_Online_Anti - Control	Dem_Online_Pro	−1.0257	−13.6203	11.5689	0.2706
Dem_Online_Anti - Control	GOP_Offline_Pro	−0.6521	−13.5225	12.2183	0.2971
Dem_Online_Anti - Control	GOP_Online_Anti	−0.8729	−13.9136	12.1678	0.3100
Dem_Online_Anti - Control	GOP_Online_Pro	−1.2288	−14.3124	11.8548	0.3110
Dem_Online_Pro - Control	Control	−2.8094	−15.7233	10.1046	0.6134
Dem_Online_Pro - Control	Dem_Offline_Pro	−2.5432	−15.6104	10.5240	0.6419
Dem_Online_Pro - Control	Dem_Online_Anti	−2.5993	−15.5335	10.3349	0.6382
Dem_Online_Pro - Control	Dem_Online_Pro	−2.8777	−15.4893	9.7339	0.6122
Dem_Online_Pro - Control	GOP_Offline_Pro	−2.6480	−15.6023	10.3062	0.6319
Dem_Online_Pro - Control	GOP_Online_Anti	−2.6063	−15.7539	10.5414	0.6325
Dem_Online_Pro - Control	GOP_Online_Pro	−2.6622	−15.8466	10.5221	0.6249
GOP_Offline_Pro - Control	Control	−1.3965	−14.3203	11.5272	0.6540
GOP_Offline_Pro - Control	Dem_Offline_Pro	−1.7841	−14.8171	11.2488	0.6361
GOP_Offline_Pro - Control	Dem_Online_Anti	−1.7823	−14.7004	11.1359	0.6328
GOP_Offline_Pro - Control	Dem_Online_Pro	−1.5053	−14.1613	11.1507	0.6403
GOP_Offline_Pro - Control	GOP_Offline_Pro	−1.6732	−14.6151	11.2688	0.6395
GOP_Offline_Pro - Control	GOP_Online_Anti	−1.5999	−14.7134	11.5136	0.6485
GOP_Offline_Pro - Control	GOP_Online_Pro	−1.4637	−14.6171	11.6896	0.6570
GOP_Online_Anti - Control	Control	−2.1253	−14.8114	10.5607	0.8086
GOP_Online_Anti - Control	Dem_Offline_Pro	−2.3659	−15.1605	10.4288	0.7914
GOP_Online_Anti - Control	Dem_Online_Anti	−2.2159	−14.9085	10.4767	0.8000
GOP_Online_Anti - Control	Dem_Online_Pro	−1.8043	−14.2558	10.6471	0.8260
GOP_Online_Anti - Control	GOP_Offline_Pro	−2.2192	−14.9297	10.4913	0.8008
GOP_Online_Anti - Control	GOP_Online_Anti	−2.4240	−15.2847	10.4367	0.7897
GOP_Online_Anti - Control	GOP_Online_Pro	−2.4414	−15.3334	10.4507	0.7900
GOP_Online_Pro - Control	Control	−2.3417	−15.1428	10.4594	0.5256
GOP_Online_Pro - Control	Dem_Offline_Pro	−1.7263	−14.6643	11.2117	0.5646
GOP_Online_Pro - Control	Dem_Online_Anti	−1.7783	−14.6150	11.0584	0.5611
GOP_Online_Pro - Control	Dem_Online_Pro	−2.2970	−14.8738	10.2799	0.5278
GOP_Online_Pro - Control	GOP_Offline_Pro	−1.9281	−14.7743	10.9181	0.5517
GOP_Online_Pro - Control	GOP_Online_Anti	−1.9621	−14.9516	11.0275	0.5500
GOP_Online_Pro - Control	GOP_Online_Pro	−2.1447	−15.1550	10.8656	0.5385
Independent - Democrat	Control	5.6630	−10.8959	22.2218	0.5027
Independent - Democrat	Dem_Offline_Pro	−11.9652	−27.8848	3.9543	0.1407
Independent - Democrat	Dem_Online_Anti	−1.9994	−18.7681	14.7692	0.8152
Independent - Democrat	Dem_Online_Pro	8.6852	−9.7946	27.1650	0.3570
Independent - Democrat	GOP_Offline_Pro	7.9771	−8.9647	24.9189	0.3561
Independent - Democrat	GOP_Online_Anti	−5.4266	−21.1483	10.2950	0.4987
Independent - Democrat	GOP_Online_Pro	6.1146	−10.0658	22.2950	0.4589
Republican - Democrat	Control	3.3775	−8.0819	14.8369	0.5635
Republican - Democrat	Dem_Offline_Pro	−1.6636	−13.3415	10.0144	0.7801
Republican - Democrat	Dem_Online_Anti	−12.4100	−24.2448	−0.5751	0.0399
Republican - Democrat	Dem_Online_Pro	−0.0049	−11.5773	11.5676	0.9993
Republican - Democrat	GOP_Offline_Pro	9.5470	−2.4004	21.4945	0.1173
Republican - Democrat	GOP_Online_Anti	4.8855	−6.8532	16.6242	0.4147
Republican - Democrat	GOP_Online_Pro	−5.6539	−17.6665	6.3587	0.3563

[1] This table presents the estimated pairwise contrasts for different treatment groups.

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 14: Pairwise contrasts for interaction terms show no significant effects for online content moderation

Pairwise Contrasts for Interaction Effects regarding Public Speech Regulation[1]

Estimated Interaction Effects, Confidence Intervals, and P-values

contrast	treat	Estimated Effect	95% CI Lower	95% CI Upper	P-value
Dem_Offline_Pro - Control	Control	1.7276	−6.4705	9.9256	0.6056
Dem_Offline_Pro - Control	Dem_Offline_Pro	1.7884	−6.4225	9.9993	0.5999
Dem_Offline_Pro - Control	Dem_Online_Anti	1.8104	−6.4241	10.0448	0.5978
Dem_Offline_Pro - Control	Dem_Online_Pro	1.9399	−6.3114	10.1913	0.5857
Dem_Offline_Pro - Control	GOP_Offline_Pro	1.9688	−6.3000	10.2375	0.5830
Dem_Offline_Pro - Control	GOP_Online_Anti	1.7517	−6.4409	9.9443	0.6034
Dem_Offline_Pro - Control	GOP_Online_Pro	1.8941	−6.3475	10.1357	0.5900
Dem_Online_Anti - Control	Control	−0.4341	−8.7415	7.8732	0.2785
Dem_Online_Anti - Control	Dem_Offline_Pro	−0.2776	−8.5975	8.0422	0.2821
Dem_Online_Anti - Control	Dem_Online_Anti	−0.2709	−8.6110	8.0693	0.2785
Dem_Online_Anti - Control	Dem_Online_Pro	0.0897	−8.2683	8.4478	0.2889
Dem_Online_Anti - Control	GOP_Offline_Pro	0.1342	−8.2393	8.5076	0.2877
Dem_Online_Anti - Control	GOP_Online_Anti	−0.3444	−8.6479	7.9591	0.2826
Dem_Online_Anti - Control	GOP_Online_Pro	−0.0280	−8.3766	8.3206	0.2862
Dem_Online_Pro - Control	Control	−0.1064	−8.3148	8.1019	0.2998
Dem_Online_Pro - Control	Dem_Offline_Pro	0.0587	−8.1577	8.2751	0.3021
Dem_Online_Pro - Control	Dem_Online_Anti	0.1091	−8.1279	8.3460	0.2939
Dem_Online_Pro - Control	Dem_Online_Pro	0.4656	−7.7788	8.7100	0.3036
Dem_Online_Pro - Control	GOP_Offline_Pro	0.5384	−7.7204	8.7972	0.2993
Dem_Online_Pro - Control	GOP_Online_Anti	−0.0359	−8.2376	8.1658	0.3057
Dem_Online_Pro - Control	GOP_Online_Pro	0.3413	−7.8970	8.5795	0.3020
GOP_Offline_Pro - Control	Control	−0.1519	−8.3414	8.0375	0.5536
GOP_Offline_Pro - Control	Dem_Offline_Pro	−0.0575	−8.2550	8.1400	0.5541
GOP_Offline_Pro - Control	Dem_Online_Anti	0.0167	−8.1994	8.2329	0.5435
GOP_Offline_Pro - Control	Dem_Online_Pro	0.1956	−8.0290	8.4203	0.5504
GOP_Offline_Pro - Control	GOP_Offline_Pro	0.2645	−7.9734	8.5024	0.5442
GOP_Offline_Pro - Control	GOP_Online_Anti	−0.1369	−8.3208	8.0470	0.5598
GOP_Offline_Pro - Control	GOP_Online_Pro	0.1244	−8.0942	8.3429	0.5501
GOP_Online_Anti - Control	Control	0.4870	−7.7820	8.7559	0.3123
GOP_Online_Anti - Control	Dem_Offline_Pro	0.5967	−7.6868	8.8802	0.3205
GOP_Online_Anti - Control	Dem_Online_Anti	0.4586	−7.8512	8.7684	0.3340
GOP_Online_Anti - Control	Dem_Online_Pro	0.7903	−7.5388	9.1194	0.3457
GOP_Online_Anti - Control	GOP_Offline_Pro	0.7359	−7.6128	9.0845	0.3558
GOP_Online_Anti - Control	GOP_Online_Anti	0.6295	−7.6334	8.8925	0.3097
GOP_Online_Anti - Control	GOP_Online_Pro	0.7082	−7.6099	9.0263	0.3394
GOP_Online_Pro - Control	Control	2.1626	−6.0844	10.4096	0.6910
GOP_Online_Pro - Control	Dem_Offline_Pro	2.2313	−6.0257	10.4883	0.6820
GOP_Online_Pro - Control	Dem_Online_Anti	2.3071	−5.9709	10.5851	0.6739
GOP_Online_Pro - Control	Dem_Online_Pro	2.4253	−5.8645	10.7150	0.6575
GOP_Online_Pro - Control	GOP_Offline_Pro	2.4884	−5.8164	10.7933	0.6502
GOP_Online_Pro - Control	GOP_Online_Anti	2.1614	−6.0799	10.4027	0.6902
GOP_Online_Pro - Control	GOP_Online_Pro	2.3733	−5.9088	10.6554	0.6642
Independent - Democrat	Control	−4.9053	−14.4050	4.5944	0.3115
Independent - Democrat	Dem_Offline_Pro	−2.9059	−12.2272	6.4155	0.5412
Independent - Democrat	Dem_Online_Anti	−2.8129	−12.3094	6.6836	0.5615
Independent - Democrat	Dem_Online_Pro	−0.0411	−9.2984	9.2161	0.9931
Independent - Democrat	GOP_Offline_Pro	0.6652	−8.2690	9.5995	0.8840
Independent - Democrat	GOP_Online_Anti	−12.2080	−21.7618	−2.6541	0.0123
Independent - Democrat	GOP_Online_Pro	0.4852	−8.7218	9.6922	0.9177
Republican - Democrat	Control	5.0167	−5.8775	15.9110	0.3668
Republican - Democrat	Dem_Offline_Pro	1.8308	−9.1069	12.7685	0.7429
Republican - Democrat	Dem_Online_Anti	−5.4635	−16.6258	5.6988	0.3374
Republican - Democrat	Dem_Online_Pro	−4.0463	−15.1993	7.1067	0.4770
Republican - Democrat	GOP_Offline_Pro	1.9247	−8.9026	12.7521	0.7275
Republican - Democrat	GOP_Online_Anti	−8.9108	−19.9047	2.0831	0.1122
Republican - Democrat	GOP_Online_Pro	3.7693	−7.1031	14.6417	0.4968

[1] This table presents the estimated pairwise contrasts for different treatment groups.

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

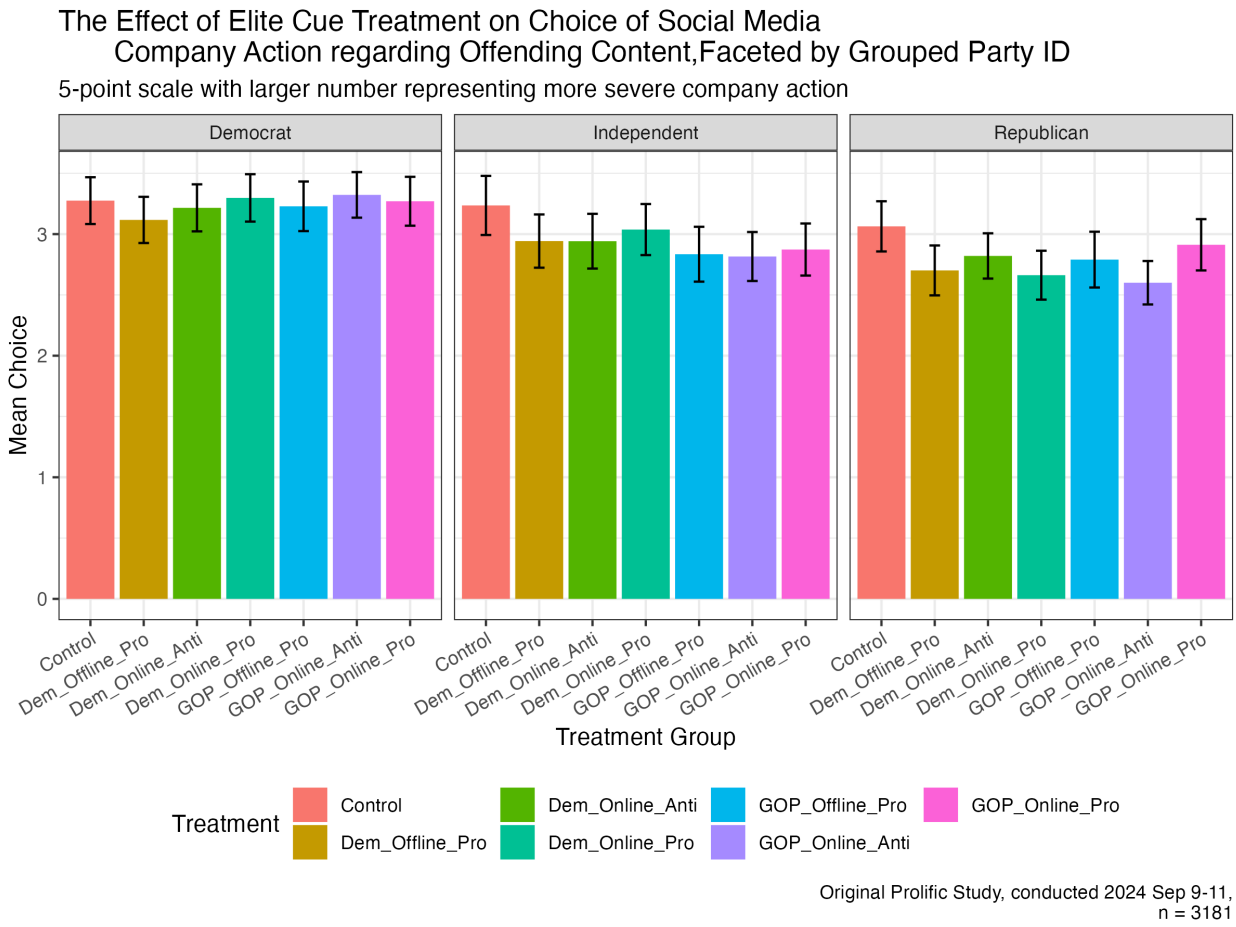
Figure 15: Pairwise contrasts for interaction terms show no significant effects for public speech regulation

oppose challenges to public free speech after exposure to cues that call for greater online content moderation. Due to participants rejecting calls to strengthen public speech laws to a greater extent than their opposition to strengthening online content moderation, I can accept hypothesis *H1*, finding that preferences for public speech regulation tended to change in a more negative direction than did changes to attitudes about online content moderation.

It should be noted that I do find some partisan heterogeneity in my effect on edge cases. Those who identify as Strong Republicans are most likely to have their attitudes significantly changed when exposed to any cue involving free speech abatement. This one subset of individuals tends to disapprove of public speech regulation so fervently that they react to any cue calling for speech regulation, whether online or offline speech and whether pro-speech regulation or anti-speech regulation, by doubling down in expressing their opposition to speech regulation of any variety.

These results from analysis of my main dependent variables persist when I conduct further analysis on my additional, secondary dependent variables. Notably though, elite cues displayed more consistent significance in altering attitudes about specific examples of law or content moderation policies that could be implemented. For instance, exposure to elite cues related to speech regulation caused Americans to generally and significantly express less support for private company intervention to moderate online content. These treatment effects dissipated in my interaction models. Still, enduring significance in that model's interaction terms suggests heterogeneous effects, with Republicans and Independents being significantly influenced to change their attitudes in the negative direction when exposed to a GOP cue calling for less content moderation. Republicans were also found to significantly oppose content moderation when exposed to a pro-content moderation tweet from Democratic elites, suggesting hostility to out-group attempts to increase speech moderation efforts.

Figure 12 shows less support for harsher company action across political identities after exposure to treatment, but comparatively larger drop-offs in support by Independents and Republicans. Still, it should be noted that most Americans, regardless of politi-


The Effect of Elite Cue Treatment on Choice of Social Media
Company Action regarding Offending Content,Faceted by Grouped Party ID
5-point scale with larger number representing more severe company action
Democrat
Independent
Republican
Mean Choice
0
1
2
3
Control
Dem_Offline_Pro
Dem_Online_Anti
Dem_Online_Pro
GOP_Offline_Pro
GOP_Online_Anti
GOP_Online_Pro
Treatment Group
Treatment
Control
Dem_Offline_Pro
Dem_Online_Anti
Dem_Online_Pro
GOP_Offline_Pro
GOP_Online_Anti
GOP_Online_Pro
Original Prolific Study, conducted 2024 Sep 9-11,
n = 3181

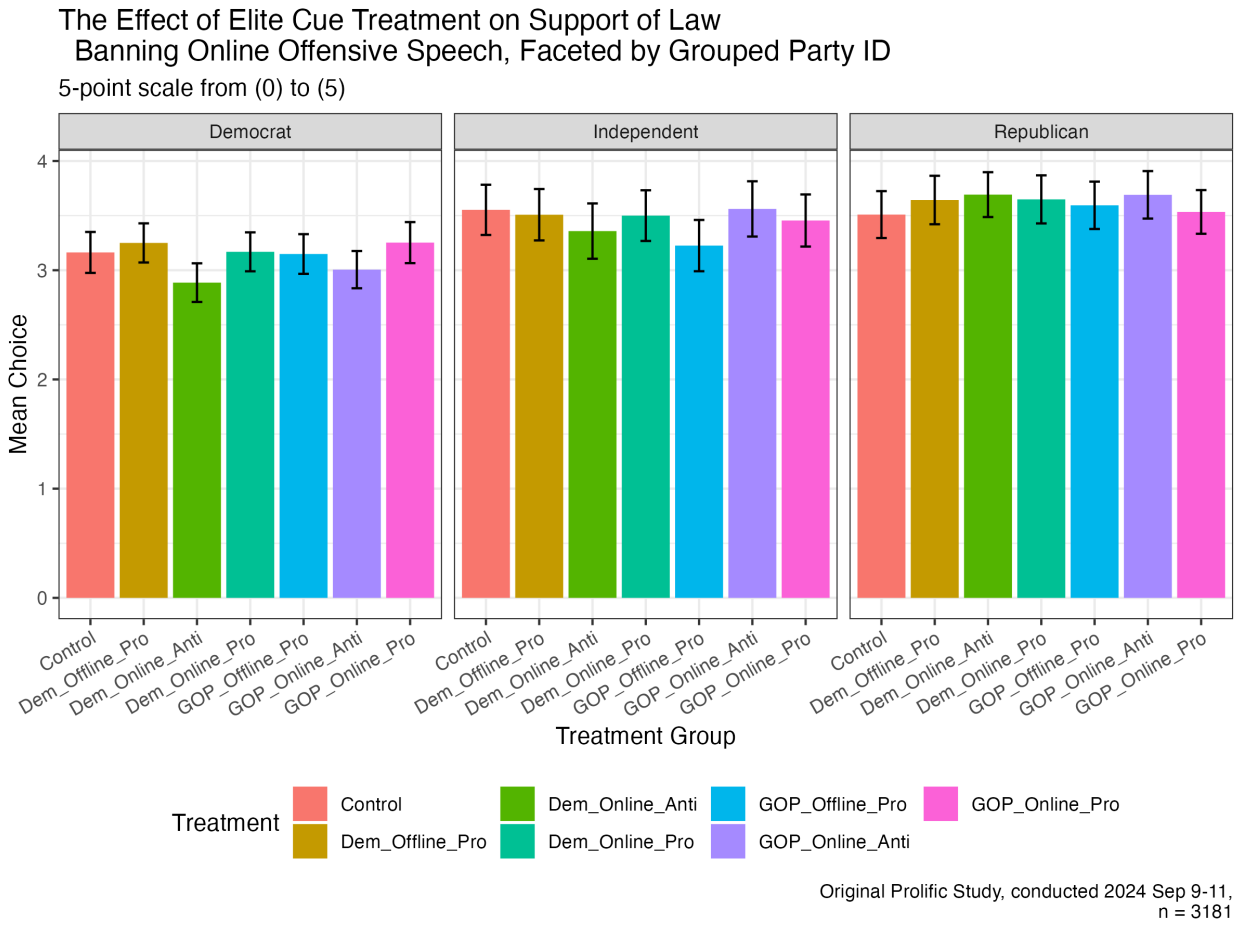


Figure 16: Preferences for Law Banning Offensive Online Speech

cal identity, tend to express their support for content moderation after exposure to a treatment as closest to a social media company policy of "Permanently remove these [offending] posts."

When analyzing dependent measures that explicitly ask about Americans' preferences for direct governmental intervention in speech, I find that exposure to elite cues has no significant effect on preferences to pass legislation that would make it illegal to post online offensive speech targeting particular groups or make it illegal to say hate speech in public against individuals or groups. The vast majority of Americans state they "Somewhat oppose" laws of these type.

The overall lack of movement in support for speech regulation after exposure to elite cues is perhaps unsurprising when analyzing how Americans feel about the trade-off

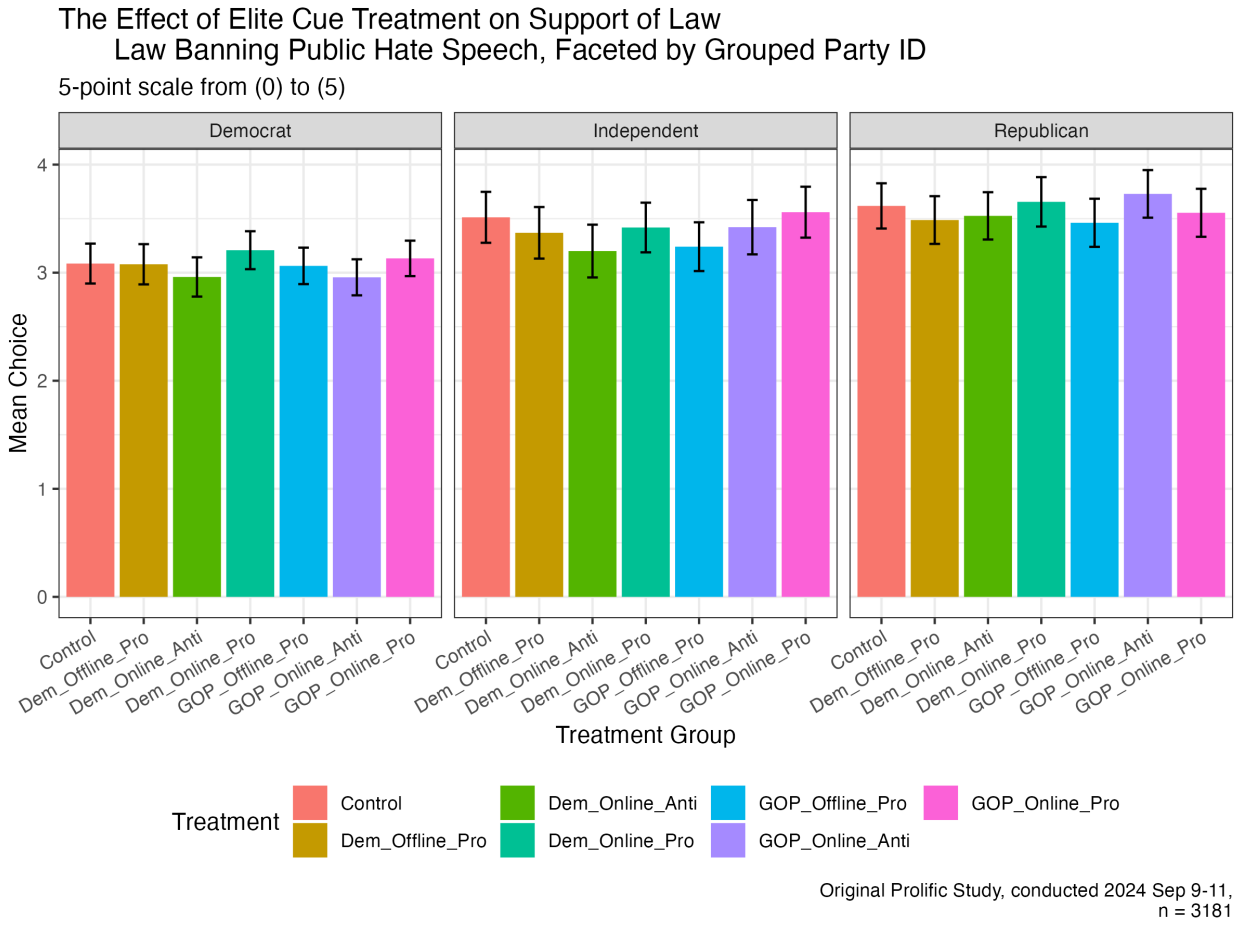


Figure 17: Preferences for Law Banning Public Hate Speech

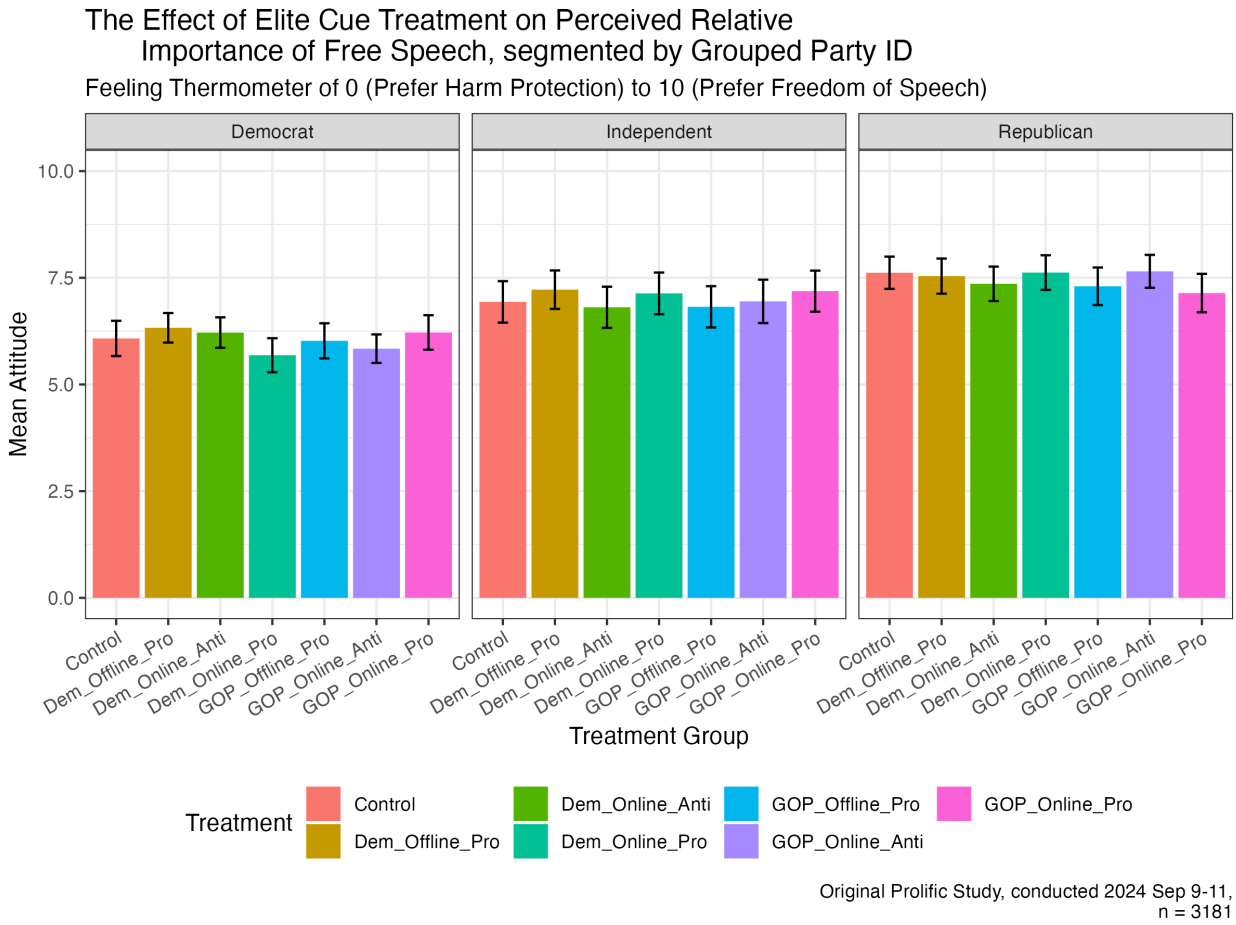


Figure 18: Attitudes about the Value of Free Speech

between freedom of speech relative to the harm it might cause. Even after exposure to elite cues urging the U.S. to adopt greater speech moderation after an assassination attempt, Democrats, Republicans, and Independents all indicate they prefer preserving freedom of speech compared to constraining speech to protect the collective from harm. Evidently, Americans greatly value free speech rights, even though Republicans report they generally value free speech relative to harm it might create to a greater extent than Democrats.

4.1 Post-Treatment post-treatment attention checks and Time Checks

Given the lack of significance of my treatments in altering attitudes, I analyzed my post-treatment attention check and time completion data to ascertain whether participants

were able to correctly identify the elite messenger who delivered their assigned cue, the content of their assigned elite cue, and whether some participants sped through my study without given adequate time to read their treatments. I find suggestive evidence that approximately 1/3 of my participants identified the source of their elite cue, but failed to identify the content of the message itself. This suggests that Americans overindex on who is saying a message versus the content of what is actually being said. This disconnect is most pronounced when an elite makes a statement that is contradicts what members of their political party have previously. Participants especially had difficulty identifying the message of an elite cue when that message went against the message their political party typically campaign on. Because participants were well aware that their assigned treatments were hypothetical and my post-treatment attention check question asked solely to identify the message, not necessarily whether they thought said message was realistic, participants should have been able to answer this post-treatment attention check correctly with accurate recall. The pilot nature of this study, however, means these findings are only suggestive of broader trends and warrant further research.

With this being said, I re-ran my analysis using filtered versions of my datasets to capture participants who (a) finished my study more than 1 standard deviation below the mean time to complete my experiment, (b) completed at least one of two post-treatment attention checks correctly, and (c) completed both post-treatment attention checks correctly. Analysis on these filtered datasets suggest that Americans who were able to identify both post-treatment attention checks correctly might be more likely to have their attitudes heterogeneously shifted after exposure to an elite cue.

In fact, regression analysis on participants ability to correctly identify the right pro or anti-speech moderation message they read in their treatments showed that there were significant differences across groups, at least compared to participants ability to correctly identify the message of the most likely elite cue of Democratic elites advocating for greater online content moderation.

Additional plots, tables, and analysis can be found in my Appendix.

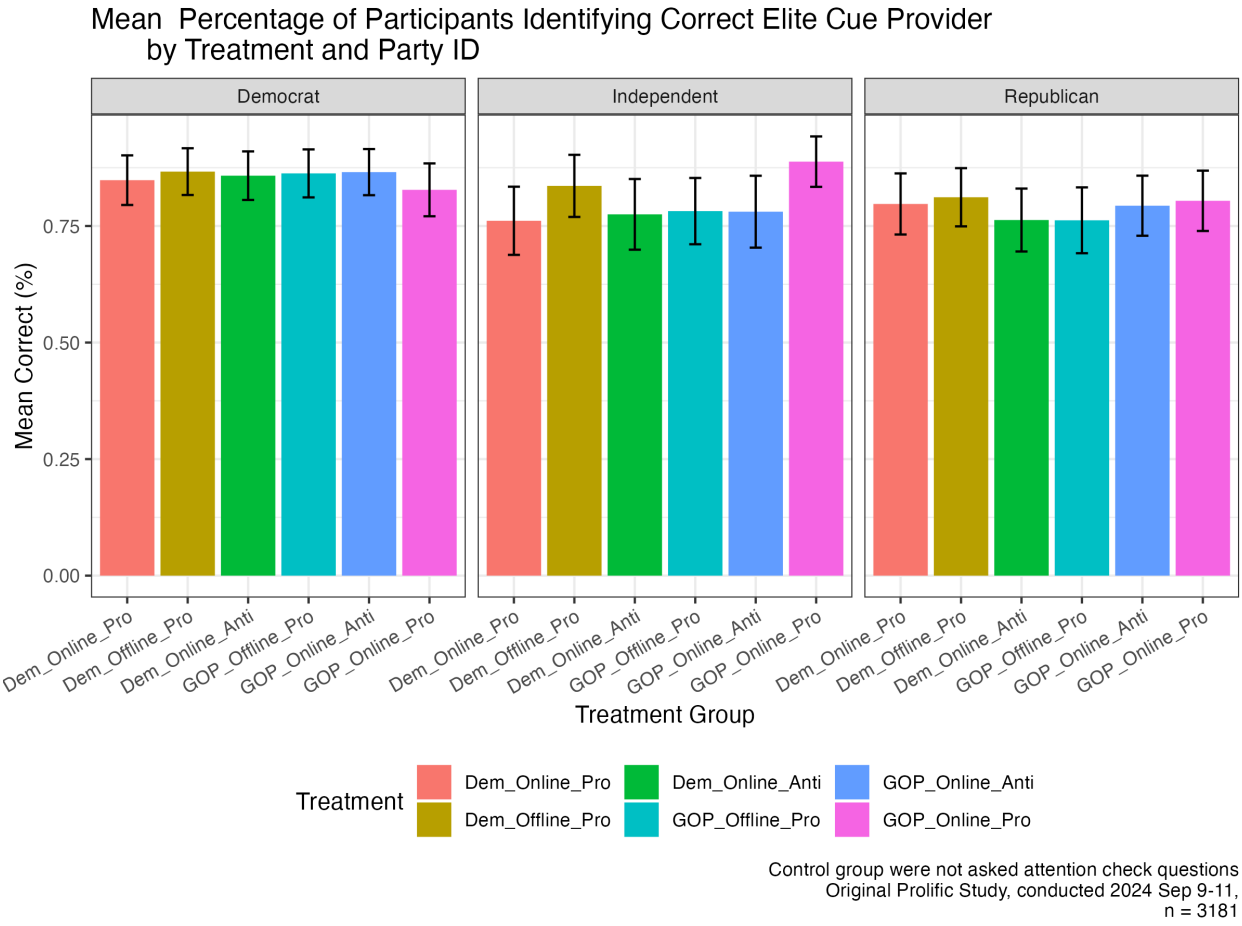


Figure 19: Ability to identify by elite cue messenger by group

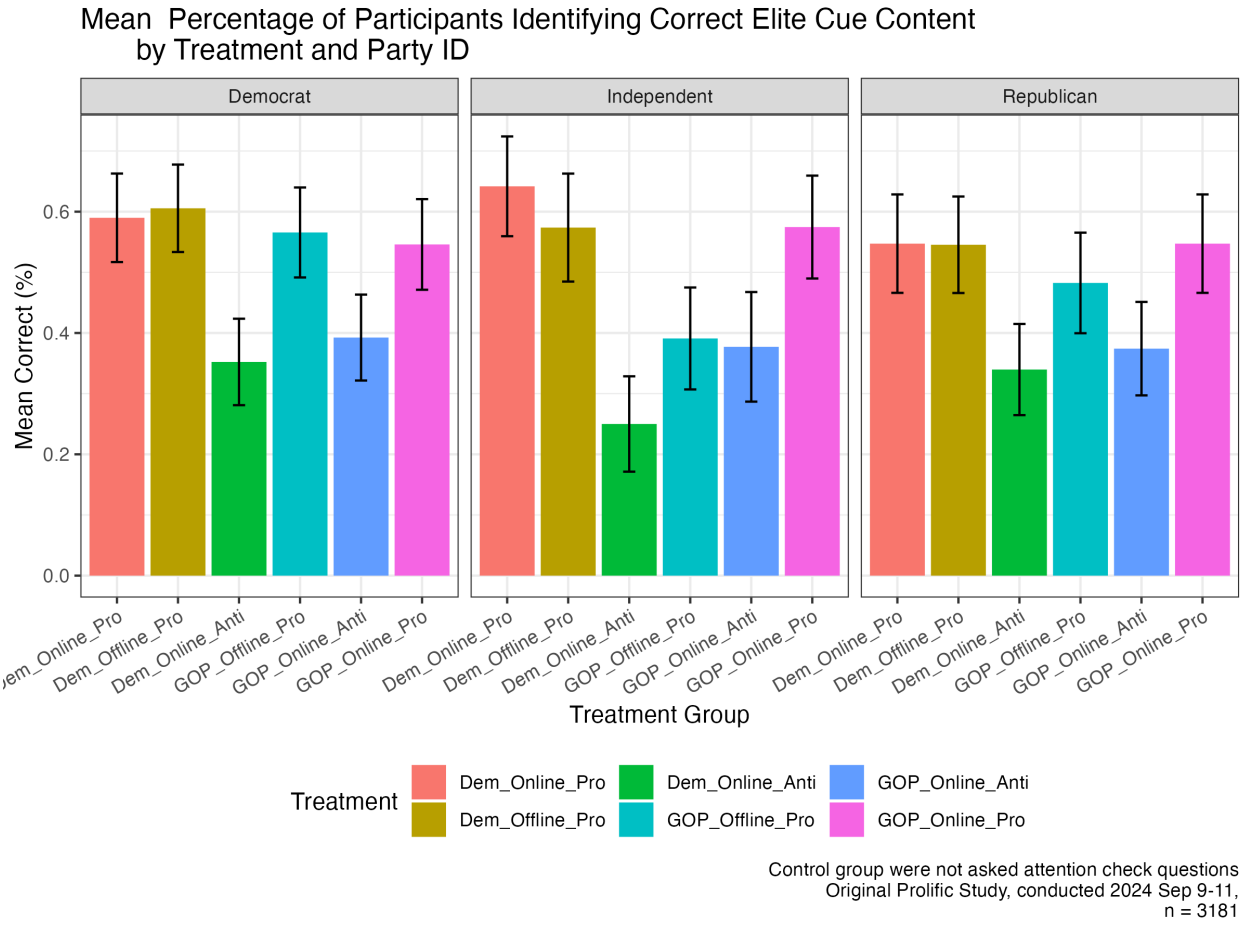


Figure 20: Ability to identify by elite cue content by group

5 Discussion

The results from this study add to the story that there might be more of a perception of there being extremely polarized disagreement in values between political identity groups over what to do about content moderation than what exists in reality. This study suggests that Americans tend to be supportive of free speech rights overall, regardless of their partisan identity or affinity to their political in-group. I do not find evidence that elite cues are powerful enough to significantly alter within partisan identity groups attitudes about either online content moderation or public speech regulation. Therefore, at least in this study's case, evidence suggests that American partisans' attitudes about content moderation are relatively durable and are not susceptible to one-off elite cues.

The design of my study cannot ascertain whether my cues were too weak or unpersuasive to convince Americans to change their policy preferences regarding content moderation, or whether Americans' interval values on speech moderation were so strong as to insulate their policy preferences from a challenge presented by an elite cue. Yet, my study does raise questions as to whether free speech is simply a bedrock American cultural value that Americans are unwilling to compromise on. For instance, my finding that Americans of all political persuasion favor freedom of speech over relative harm from speech after exposure to elite cues, and that Americans tend to indicate less support for speech moderation after exposure to elite cues calling for moderation indicates that Americans very strongly value freedom of speech, or at least an American conception of freedom of speech. That is to say, my study suggests that Americans might very well posses a cultural conception of a uniquely version of free speech influenced by American legal and political history. This finding comports with legal research that theorizes that the importance of the American legal system and its interest in free speech rights has created a cultural value of protecting free speech within the United States that is exceptional compared to other democratic countries (Krotoszynski, 2015; Alvarez and Kemmelmeier, 2017). My results finding that most Americans agree that free speech rights should be prioritized over increased speech moderation might very well differ from what citizens of other countries would indicate if they were to take this

experiment. Polling research, for instance, shows that Americans are generally more supportive of freedom of expression in all forms, including calls for violent protest and offensive statements, than others worldwide (Simmons, 2015; Staghøj, 2021).

This study also finds evidence that unity messaging in the wake of attempted political violence might not have a significant effect in garnering support for codified content moderation. Such a finding raises interesting questions for real-world politics in an era of political polarization. While my study cannot conclude whether negatively-framed elite cues might have been more effective in altering attitudes about speech moderation, it can confidently conclude that one-off elite cues of the nature this study used as a treatment are not effective.

That is not to say there still is not scope to theorize about the role of elite cues in shifting attitudes about free speech. For instance, consistent, well-managed messaging campaigns from political elites might very well be successful in altering policy preferences for online content moderation. There is past precedent for these sorts of consistent information claims in democratic countries across the world, of course. Yet, what my study finds is that one-off elite cues, such as a tweet that might not be consistent with past statements from a politician, are not effective in altering attitudes. This finding is perhaps raising a larger question about the role of social media and X (Twitter) itself in modern political society. Perhaps Americans have become so desensitized to content online, even tweets from official political accounts, that they take everything they read online with a grain of salt. In other words, it is possible that Americans simply discount the elite cues they read online. In such a world, media concern about single tweets or elite statements altering the political landscape would be overblown. A tweet that seems real but odd might not cause a red-button detonator event as Americans would have learned that not everything said online is meant to be taken seriously, even by politicians.

My findings also continue to raise important questions about the practical impacts brought by political polarization in society. I have found evidence that Americans have different preferences regarding the appropriate level of content moderation in society divided by party lines. However, most of my results indicated that in a broader sense

Americans did not disagree whether or not there should be any content moderation, just how much. There may indeed be opportunities to find common ground between Republicans and Democrats, particularly as both groups are largely supportive of free speech rights in the U.S. overall.

Where I did find significant division, though, is in how significantly sociodemographic variables predicted attitudes about speech regulation. In particular, Black, female, and older Americans–groups that may be at greater risk of harm from harassment via speech–preferred greater online and offline speech moderation compared to non-Black, male, and younger Americans.

My study also raises an interesting question about how Americans process text-based elite cues themselves. Although my post-treatment attention check analysis results cannot be seen as definitive due to design constraints, I found that approximately 40 percent (or 1077 of 2723) of participants assigned to a treatment group were only able to identify who delivered an elite cue but failed to identify the actual message of this cue. When I restrict my sample to the 1159 participants plus control group participants who answered both post-treatment attention checks correctly, I find evidence that is more suggestive of heterogeneous partisan effects from exposure to partisan elite cues. This insight suggests that some of my participants, even though they took sufficient time to process my treatments, may have unknowingly rejected or ignored the message shared by a partisan elite when that message did not align with their expectation of what that partisan elite would support.

While this study has hopefully provided insight into the role of elite cues in affecting Americans' preferences for law and policies guiding content moderation, I am well aware that my study suffers from certain limitations. For instance, if I were to run my study again, I likely would not have included a pre-treatment measure of my main feeling thermometer outcome variables as I believe doing so might have anchored my participants before treatment. Additionally, I would consider utilizing deception in my study rather than telling my participants they were looking at hypothetical tweets, although I prefer not to employ deception in my research. If a political elite were to make a direct

appeal to increase or decrease online content moderation, future researchers should also consider employing that statement as a treatment. Unfortunately, my research did not reveal a suitable treatment from political elites that I believed was salient enough to elicit responses from participants.

There are, however, positives from my study design. Because I used a representative sample from Prolific, I am confident that my study is well-balanced. My post-treatment attention checks and pilot study revealed that the vast majority of my participants recruited on Prolific took my study seriously and understood my questions. My findings also should shed light on how Americans think about freedom of speech.

Finally, my study raises questions about the path forward in determining appropriate content moderation procedures for large scale, and increasingly globalized, social media platforms. This study suggests that Americans might *prima facie* prefer a minimalist approach to online content moderation, but such an approach would likely be insufficient in stopping toxic speech and hate speech from appearing online (Pradel et al., 2023). Such a minimal content moderation policy might also prove unacceptable to consumers in other countries that use American internet products, leading to further controversy. I believe this study's findings reinforce that future research should explore (1) whether Americans empirically value free speech differently than do citizens of other countries with strong free rights as established via law, and (2) if so, what exactly makes the American conception of free speech different than that held by others around the world. Answering these questions will reveal important insights about the increasingly relevant topic of speech governance in democratic systems.

6 Appendix

My experiment was conducted according to the University of Oxford's policy for human subjects research and was approved by the Department of Politics and International Relations CUREC committee. I gathered informed consent from each participant at the beginning of the survey and debriefed participants at the conclusion of my survey.

6.1 Power Calculations

In my in-group elite cues experiment, I calculated I required 1,190 individuals in 6 treatment groups and 1 control group for each of my political identities, Democrat, Republican, and Independent.

I ran power calculations by drawing outcomes from a normal distribution:

$$Yi \approx norm.(n)$$

where the underlying probability of getting supporting more stringent online content regulatory policy is simply modeled using the identity link function as,

$$P(Outcome)i = \beta 0 + \beta 2aT2ai + \beta 2bT2bi + \beta 2cT2ci + \beta 2dT2di + \beta 2eT2ei + \beta 2fT2fi$$

where:

- i index individuals
- $P(Outcome)i > 50$ on a 0 to 100 feeling thermometer scale, where 0 indicates complete lack of support of online content regulation and 100 indicates complete support for online content regulation, if the components sum to greater than 50 and $P(Outcome)i < 50$ when the components sum to less than 50
- $\beta 0$ is the placebo probability of supporting online content moderation following exposure to the control group, otherwise understood as the Bayesian prior preference for online content moderation when can be decomposed as the function of internal values regarding this policy issue of regulation exogenous of political cues or socialization added to the error term describing endogenous political-cultural socialization
- $\beta 2a$ is the treatment effects of being exposed to an IN-group elite cue advocating FOR content moderation in the arena of ONLINE content.
- $\beta 2b$ is the treatment effects of being exposed to an IN-group elite cue advocating

AGAINST content moderation in the arena of ONLINE content.

- $\beta 2c$ is the treatment effects of being exposed to an IN-group elite cue advocating FOR content moderation in the arena of OFFLINE (in-person) content.
- $\beta 2d$ is the treatment effects of being exposed to an IN-group elite cue advocating AGAINST content moderation in the arena of OFFLINE (in-person) content.
- $\beta 2e$ is the treatment effects of being exposed to an OUT-group elite cue advocating FOR content moderation in the arena of ONLINE content.
- $\beta 2f$ is the treatment effects of being exposed to an OUT-group elite cue advocating AGAINST content moderation in the arena of ONLINE content.

I noted my hypotheses earlier during my earlier discussion of hypotheses. I focus my attention now on explaining my coefficients for these different treatments and justifying my effect sizes. The coefficient for my in-group elite cues regarding online content moderation treatments ($\beta 2a$ and $\beta 2b$) are expected to be my largest and, thus, the simplest to statistically distinguish from the placebo group. This is because I expect participants to have the most susceptible to cue exposure internal values on online content moderation, given the novelty of online communication, the recognizability of the cheapness of online speech, and the widespread publication of potential dangers that might arise from online communication. Because I hypothesized that solely being exposed to an in-group elite cue would be enough to influence in-group partisans to adopt stated policy preferences different from their original attitude about online content moderation, I expect my effect sizes for this treatment to be sufficiently large as to "flip" participants who express weak a priori attitudes about the topic. To recover an estimate of the minimum effect size for the in-group only treatment effect ($T2a$ and $T2b$), with 0.8 power at the 95% confidence level, I begin by assuming $\beta 2a = 0.2$ and $\beta 2b = -0.2$.

The coefficient for my in-group elite cues regarding offline content moderation treatments ($\beta 2c$ and $\beta 2d$) are expected to be my smallest and, thus, the most difficult to statistically distinguish from the placebo group. This is because I expect participants to have the least susceptible to cues internal values on offline content moderation, given

how entrenched free speech rights are in the US legal, constitutional, and social contexts. Yet, because I hypothesized that solely being exposed to an in-group elite cue will have a strong effect on even strongly held Bayesian priors, I expect my effect sizes for this treatment to be noticeable although only so large as to "flip" participants who express extremely weak (or moderate and central) a priori attitudes about the topic. To recover an estimate of the minimum effect size for the in-group only treatment effect ($T2c$ and $T2d$), with 0.8 power at the 95% confidence level, I begin by assuming $\beta 2c = 0.05$ and $\beta 2d = -0.05$.

The coefficient for my out-group elite cues regarding online content moderation treatments ($\beta 2e$ and $\beta 2f$) are expected to be in between the two previous sets of treatment arms. This is because I expect participants to react in the opposite direction as they usually would when exposed to an outgroup, or antagonistic elite cue. That is because I hypothesized that solely being exposed to an out-group elite cue would have an effect on Bayesian priors, either entrenching those priors if the out-group indicates support for an opposing stance on regulation or forcing the abandonment of priors if the outgroup signals support for these priors, I expect my effect sizes for this treatment to be noticeable and in between the two other pairs of treatment arms. To recover an estimate of the minimum effect size for the in-group-only treatment effect ($T2e$ and $T2f$), with 0.8 power at the 95% confidence level, I begin by assuming $\beta 2e = -0.1$ and $\beta 2f = 0.1$.

I then generated 5000 hypothetical datasets, each set of 1000 of these datasets set to either 140, 560, 1050, 1190, or 2380 subjects. I assumed equal probability of assignment to the six treatment arms plus one control group. For each simulated dataset, I realized the outcome for each subject and then fit the model in my equation above.

I estimated the power of the simulated experiment as the proportion of Outcome coefficients ($\beta 2a$, $\beta 2b$, $\beta 2c$, $\beta 2d$, $\beta 2e$, $\beta 2f$) across these 1000 models per each n-size with p-values below 0.05 (using the conventional two-sided test). If this power is above 0.8, I decreased the hypothesized effect size by 5% and repeated the simulation, until the estimated power fell below 0.8. The final well-powered effect size was my estimate of the minimum effect size we can reliably detect.

I also added covariates to my model, but find only minimal improvements to power through the addition of an ideology covariate. Therefore, I do not include this covariate power curve or model in the figures below.

Figure 3 reports these minimum effect size estimates as power curves. If my effect sizes are as hypothesized in the scenario described in this section, I expect to be well-powered to observe an effect size of 0.2 and above 3 for my $\beta 2a$ treatment arm in one of my blocked experiments with 1190 participants. If my effect size as described above is underestimated and effect sizes are in fact 1.5x than as reported in this section, I expect to be well-powered to observe an effect size of .3 for my $\beta 2a$ treatment arm with 550 participants. If my effect size as described above is overestimated and effect sizes are in fact .75x as reported in this section, I expect to be well-powered to observe an effect size of .15 for my $\beta 2a$ treatment arm with (1190 times 3 political identity groups) = 3570 participants. However, as I was able to use a representative sample via Proflific due to free representative sampling being available to academic researchers, I require fewer participants and estimate instead that I only require 3300 participants versus the calculated 3570.

The simulations are conducted in R. The graph of my power curves for the effect size scenarios described in this section can be seen below.

6.2 Further Analyses

6.3 Survey Experiment

Start of Block: Consent

Consent Form

Consent to Participate in a Research Study**University of Oxford**

Consent to Participate in a Research Study

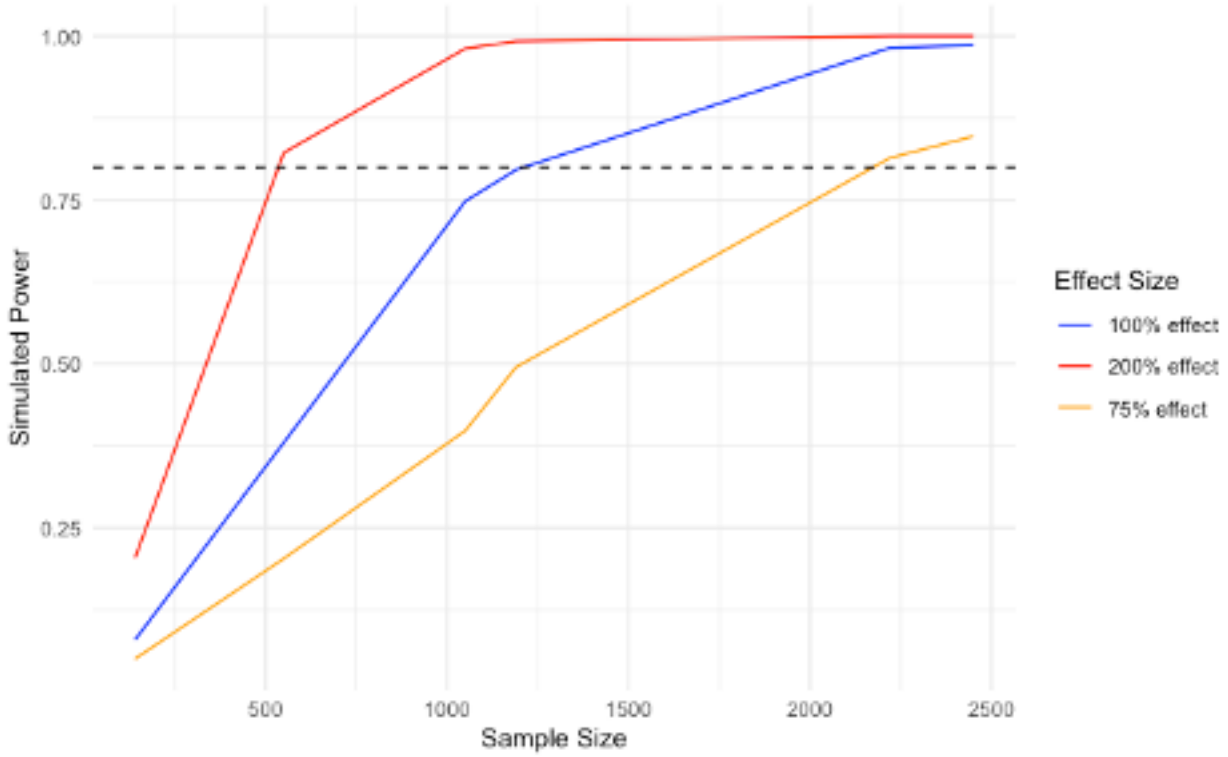


Figure 21: Generated power curves for participants needed for each Political ID

Welcome to our study entitled "Public Preferences on Social Values"! This is a study examining attitudes and political views and has received ethical approval from the University of Oxford (SSH_DPIR_C1A_24_041). We are generally interested in understanding how people make up their minds and how different political debates on social media elicit different responses. The principal investigator of this study is Dr. X (University of Oxford) and the main researcher is candidate X (University of Oxford). Feel free to contact us through the Centre for Experimental Social Science at cess@nuffield.ox.ac.uk if you have any questions, to provide feedback about your experience or if you are interested in the goals of this research (or through our personal emails XXXX OR XXXX).

This consent form asks you to allow the researcher to record your responses and to use the anonymized data to further understand social and political values citizens might have. We will ask you questions about your attitudes and views. We will also show you some hypothetical, generated screenshots of political statements based on

Comparison of Difference-in-Difference Regression Models of Elite Cue Treatment Group Against Favorability Towards Online Content Moderation

Difference of Pre-Treatment Feeling Thermometer Subtracted from Post-Treatment Feeling Thermometer (More positive numbers indicate increase in support)

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	6.263***	14.46**	4.253**	11.88*
treat				
Control	—	—	—	—
Dem_Offline_Pro	0.6939	0.0347	4.219	4.301
Dem_Online_Anti	-1.419	-1.018	-1.048	-1.099
Dem_Online_Pro	-1.222	-0.5451	1.528	2.731
GOP_Offline_Pro	-0.8609	-1.792	0.9358	0.8450
GOP_Online_Anti	-0.5961	-0.5372	3.011	4.806
GOP_Online_Pro	0.8030	0.9193	2.713	3.053
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-1.713	-4.435*	0.5114	-2.433
Republican	-1.801*	-7.313*	2.320	-1.923
Party_Affinity		-1.634*		-1.647*
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		-1.838		-1.792
Gender				
Female		—		—
Male		0.8947		0.7784
Non-binary		2.516		3.098
Ethnicity				
White		—		—
Asian		-0.3849		-0.2833
Black		2.350		2.402
Latino		3.391		3.630
Multiracial		0.6693		0.9279
Native American		2.573		1.943
Age		0.5384		0.5382
Socials				
At least once a week but not every day		—		—
A few times a month		1.981		1.964
Every day		-2.042		-1.903
Less often or not at all		-4.353		-4.180
Community				
Suburb		—		—
City		0.9996		0.9727
Other		-6.228		-4.924
Rural		-0.1593		0.0188
Town		2.776		2.761
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		0.3028		0.1795
Graduate degree		-0.2398		-0.2339
Prefer not to say		-2.022		-1.254
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			-5.467	-4.774
Dem_Online_Anti * Independent			-0.2243	2.831
Dem_Online_Pro * Independent			-2.904	-3.387
GOP_Offline_Pro * Independent			0.8564	0.2344
GOP_Online_Anti * Independent			-6.108	-6.562
GOP_Online_Pro * Independent			-2.029	-2.889
Dem_Offline_Pro * Republican			-6.039	-8.296*
Dem_Online_Anti * Republican			-0.9096	-1.906
Dem_Online_Pro * Republican			-5.804	-6.534
GOP_Offline_Pro * Republican			-6.313	-7.868
GOP_Online_Anti * Republican			-5.952	-10.08*
GOP_Online_Pro * Republican			-3.989	-3.491

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 22: Comparison of Pre-Post Post-Treatment Online Content Moderation Preferences

Comparison of Difference-in-Difference Regression Models of Elite Cue Treatment Group Against Favorability Towards Public Speech Regulation

Difference of Pre-Treatment Feeling Thermometer Subtracted from Post-Treatment Feeling Thermometer (More positive numbers indicate increase in support)

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	6.263***	4.387	4.253**	1.181
treat				
Control	—	—	—	—
Dem_Offline_Pro	0.6939	-1.774	4.219	-2.195
Dem_Online_Anti	-1.419	-2.568	-1.048	1.519
Dem_Online_Pro	-1.222	-0.6717	1.528	3.000
GOP_Offline_Pro	-0.8609	-1.985	0.9358	1.965
GOP_Online_Anti	-0.5961	-0.2528	3.011	4.732
GOP_Online_Pro	0.8030	0.2013	2.713	3.886
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-1.713	-3.228	0.5114	-0.4643
Republican	-1.801*	-4.269	2.320	1.587
Party_Affinity		-0.6789		-0.6591
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		-0.5074		-0.4891
Gender				
Female		—		—
Male		-0.2420		-0.1551
Non-binary		-0.5711		-0.1617
Ethnicity				
White		—		—
Asian		2.745		2.882
Black		3.761*		3.816**
Latino		0.6237		1.105
Multiracial		0.9699		1.073
Native American		-4.447		-4.580
Age		-0.0087		0.0267
Socials				
At least once a week but not every day		—		—
A few times a month		-1.511		-1.434
Every day		-1.328		-1.231
Less often or not at all		-0.8212		-0.7677
Community				
Suburb		—		—
City		-0.4288		-0.4884
Other		0.8438		2.290
Rural		1.306		1.381
Town		1.573		1.466
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		-1.094		-1.224
Graduate degree		-0.1475		-0.3031
Prefer not to say		11.26		11.01
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			-5.467	3.314
Dem_Online_Anti * Independent			-0.2243	-5.056
Dem_Online_Pro * Independent			-2.904	-3.459
GOP_Offline_Pro * Independent			0.8564	-3.250
GOP_Online_Anti * Independent			-6.108	-5.845
GOP_Online_Pro * Independent			-2.029	-5.015
Dem_Offline_Pro * Republican			-6.039	-1.096
Dem_Online_Anti * Republican			-0.9096	-7.434
Dem_Online_Pro * Republican			-5.804	-7.775*
GOP_Offline_Pro * Republican			-6.313	-8.753*
GOP_Online_Anti * Republican			-5.952	-9.593*
GOP_Online_Pro * Republican			-3.989	-6.348

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 23: Comparison of Pre-Post Post-Treatment Online Content Moderation Preferences

Comparison of Regression Models of Elite Cue Treatment Group Against Appropriate Social Media Company Action regarding Guideline-violating Content

More positive numbers indicate harsher company actions

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	3.427***	3.013***	3.275***	2.706***
treat				
Control	—	—	—	—
Dem_Offline_Pro	-0.2649**	-0.2490*	-0.1586	0.0170
Dem_Online_Anti	-0.1853*	-0.1842	-0.0594	0.0682
Dem_Online_Pro	-0.1776*	-0.2123*	0.0225	0.1466
GOP_Offline_Pro	-0.2227**	-0.3231**	-0.0467	-0.0498
GOP_Online_Anti	-0.2489**	-0.2638*	0.0473	0.2863
GOP_Online_Pro	-0.1557	-0.0825	-0.0052	0.1514
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-0.2936***	-0.0557	-0.0395	0.3674
Republican	-0.4547***	-0.1289	-0.2116	0.3259
Party_Affinity		0.0282		0.0289
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		-0.3563***		-0.3460***
Gender				
Female		—		—
Male		0.0487		0.0523
Non-binary		0.1824		0.1824
Ethnicity				
White		—		—
Asian		0.1203		0.1320
Black		0.3612***		0.3614***
Latino		0.0483		0.0789
Multiracial		-0.1798		-0.1574
Native American		-0.2196		-0.2304
Age		0.0465*		0.0503**
Socials				
At least once a week but not every day		—		—
A few times a month		0.0595		0.0601
Every day		-0.0284		-0.0212
Less often or not at all		0.2962*		0.3095*
Community				
Suburb		—		—
City		0.1190		0.1156
Other		-1.670		-1.591
Rural		0.0916		0.0951
Town		-0.0272		-0.0340
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		-0.1334		-0.1386
Graduate degree		-0.0019		-0.0150
Prefer not to say		1.419		1.501
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			-0.1345	-0.4613
Dem_Online_Anti * Independent			-0.2347	-0.3990
Dem_Online_Pro * Independent			-0.2209	-0.3805
GOP_Offline_Pro * Independent			-0.3545	-0.4823
GOP_Online_Anti * Independent			-0.4673*	-0.8056**
GOP_Online_Pro * Independent			-0.3575	-0.4419
Dem_Offline_Pro * Republican			-0.2038	-0.3975
Dem_Online_Anti * Republican			-0.1838	-0.4079
Dem_Online_Pro * Republican			-0.4240*	-0.7549**
GOP_Offline_Pro * Republican			-0.2268	-0.3909
GOP_Online_Anti * Republican			-0.5110**	-0.9532***
GOP_Online_Pro * Republican			-0.1464	-0.3147

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Pairwise Contrasts for Interaction Effects regarding Company Responsibility to Moderate Online Content[1]

Estimated Interaction Effects, Confidence Intervals, and P-values

contrast	treat	Estimated Effect	95% CI Lower	95% CI Upper	P-value
Dem_Offline_Pro - Control	Control	−0.2811	−0.5645	0.0024	0.0867
Dem_Offline_Pro - Control	Dem_Offline_Pro	−0.2799	−0.5633	0.0034	0.0876
Dem_Offline_Pro - Control	Dem_Online_Anti	−0.2812	−0.5645	0.0022	0.0866
Dem_Offline_Pro - Control	Dem_Online_Pro	−0.2794	−0.5638	0.0049	0.0878
Dem_Offline_Pro - Control	GOP_Offline_Pro	−0.2789	−0.5633	0.0055	0.0882
Dem_Offline_Pro - Control	GOP_Online_Anti	−0.2787	−0.5611	0.0038	0.0888
Dem_Offline_Pro - Control	GOP_Online_Pro	−0.2803	−0.5648	0.0043	0.0871
Dem_Online_Anti - Control	Control	−0.1974	−0.4815	0.0868	0.2701
Dem_Online_Anti - Control	Dem_Offline_Pro	−0.1962	−0.4803	0.0878	0.2731
Dem_Online_Anti - Control	Dem_Online_Anti	−0.1970	−0.4811	0.0870	0.2705
Dem_Online_Anti - Control	Dem_Online_Pro	−0.1992	−0.4842	0.0859	0.2686
Dem_Online_Anti - Control	GOP_Offline_Pro	−0.1992	−0.4844	0.0860	0.2690
Dem_Online_Anti - Control	GOP_Online_Anti	−0.1927	−0.4758	0.0905	0.2807
Dem_Online_Anti - Control	GOP_Online_Pro	−0.2003	−0.4855	0.0850	0.2658
Dem_Online_Pro - Control	Control	−0.1980	−0.4807	0.0847	0.4297
Dem_Online_Pro - Control	Dem_Offline_Pro	−0.1957	−0.4782	0.0868	0.4355
Dem_Online_Pro - Control	Dem_Online_Anti	−0.1983	−0.4809	0.0843	0.4299
Dem_Online_Pro - Control	Dem_Online_Pro	−0.1942	−0.4775	0.0891	0.4322
Dem_Online_Pro - Control	GOP_Offline_Pro	−0.1930	−0.4764	0.0903	0.4338
Dem_Online_Pro - Control	GOP_Online_Anti	−0.1934	−0.4752	0.0884	0.4461
Dem_Online_Pro - Control	GOP_Online_Pro	−0.1958	−0.4793	0.0876	0.4274
GOP_Offline_Pro - Control	Control	−0.2392	−0.5236	0.0453	0.2773
GOP_Offline_Pro - Control	Dem_Offline_Pro	−0.2375	−0.5218	0.0468	0.2812
GOP_Offline_Pro - Control	Dem_Online_Anti	−0.2387	−0.5231	0.0456	0.2778
GOP_Offline_Pro - Control	Dem_Online_Pro	−0.2415	−0.5266	0.0435	0.2757
GOP_Offline_Pro - Control	GOP_Offline_Pro	−0.2416	−0.5267	0.0435	0.2764
GOP_Offline_Pro - Control	GOP_Online_Anti	−0.2323	−0.5159	0.0512	0.2907
GOP_Offline_Pro - Control	GOP_Online_Pro	−0.2432	−0.5284	0.0420	0.2722
GOP_Online_Anti - Control	Control	−0.2782	−0.5622	0.0058	0.3467
GOP_Online_Anti - Control	Dem_Offline_Pro	−0.2752	−0.5590	0.0087	0.3521
GOP_Online_Anti - Control	Dem_Online_Anti	−0.2779	−0.5617	0.0060	0.3473
GOP_Online_Anti - Control	Dem_Online_Pro	−0.2784	−0.5634	0.0066	0.3453
GOP_Online_Anti - Control	GOP_Offline_Pro	−0.2778	−0.5629	0.0074	0.3463
GOP_Online_Anti - Control	GOP_Online_Anti	−0.2686	−0.5514	0.0142	0.3644
GOP_Online_Anti - Control	GOP_Online_Pro	−0.2810	−0.5662	0.0042	0.3406
GOP_Online_Pro - Control	Control	−0.1691	−0.4525	0.1144	0.3493
GOP_Online_Pro - Control	Dem_Offline_Pro	−0.1682	−0.4515	0.1152	0.3523
GOP_Online_Pro - Control	Dem_Online_Anti	−0.1685	−0.4519	0.1148	0.3502
GOP_Online_Pro - Control	Dem_Online_Pro	−0.1724	−0.4565	0.1117	0.3447
GOP_Online_Pro - Control	GOP_Offline_Pro	−0.1728	−0.4569	0.1114	0.3446
GOP_Online_Pro - Control	GOP_Online_Anti	−0.1641	−0.4467	0.1186	0.3618
GOP_Online_Pro - Control	GOP_Online_Pro	−0.1734	−0.4577	0.1108	0.3417
Independent - Democrat	Control	0.1543	−0.1688	0.4774	0.3493
Independent - Democrat	Dem_Offline_Pro	0.0444	−0.2766	0.3654	0.7864
Independent - Democrat	Dem_Online_Anti	−0.0849	−0.4080	0.2382	0.6065
Independent - Democrat	Dem_Online_Pro	−0.0520	−0.3688	0.2648	0.7475
Independent - Democrat	GOP_Offline_Pro	−0.1936	−0.5091	0.1218	0.2290
Independent - Democrat	GOP_Online_Anti	−0.3162	−0.6426	0.0102	0.0576
Independent - Democrat	GOP_Online_Pro	−0.1087	−0.4240	0.2066	0.4993
Republican - Democrat	Control	0.1210	−0.2577	0.4997	0.5312
Republican - Democrat	Dem_Offline_Pro	−0.0754	−0.4548	0.3040	0.6970
Republican - Democrat	Dem_Online_Anti	−0.0519	−0.4321	0.3283	0.7889
Republican - Democrat	Dem_Online_Pro	−0.2827	−0.6644	0.0990	0.1467
Republican - Democrat	GOP_Offline_Pro	−0.1420	−0.5209	0.2369	0.4628
Republican - Democrat	GOP_Online_Anti	−0.3768	−0.7597	0.0061	0.0537
Republican - Democrat	GOP_Online_Pro	−0.0042	−0.3802	0.3718	0.9825

[1] This table presents the estimated pairwise contrasts for different interaction terms of party ID*treatment groups.

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 24: Mean marginal effects for interaction terms of company responsibility to moderate

Comparison of Regression Models of Elite Cue Treatment Group Against Support of Law Banning Online Offensive Speech

More positive numbers indicate support for law

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	3.406***	3.058***	3.343***	2.993***
treat				
Control	—	—	—	—
Dem_Offline_Pro	0.0395	0.0847	0.1462	0.2163
Dem_Online_Anti	-0.0472	0.0779	-0.0302	0.1740
Dem_Online_Pro	0.0366	0.0986	0.1067	0.0951
GOP_Offline_Pro	0.0726	0.0865	0.1144	0.1221
GOP_Online_Anti	0.0030	0.0167	0.1842	0.1803
GOP_Online_Pro	0.0207	0.1262	0.0366	0.0950
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-0.4350***	-0.0939	-0.3508*	-0.0658
Republican	-0.4603***	-0.0190	-0.3427**	0.1081
Party_Affinity		0.0033		0.0031
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		-0.7414***		-0.7353***
Gender				
Female		—		—
Male		-0.3201***		-0.3199***
Non-binary		-0.5284*		-0.5389*
Ethnicity				
White		—		—
Asian		0.1190		0.1266
Black		0.3179***		0.3218***
Latino		0.1723		0.1832
Multiracial		-0.0922		-0.0836
Native American		0.4550		0.4358
Age		0.0730***		0.0740***
Socials				
At least once a week but not every day		—		—
A few times a month		0.1587		0.1623
Every day		0.0514		0.0566
Less often or not at all		0.1353		0.1395
Community				
Suburb		—		—
City		0.0501		0.0485
Other		0.6068		0.6556
Rural		0.0265		0.0245
Town		0.0935		0.0934
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		-0.0297		-0.0290
Graduate degree		0.1656**		0.1633**
Prefer not to say		0.7160		0.7999
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			-0.0971	-0.1242
Dem_Online_Anti * Independent			0.0050	-0.0736
Dem_Online_Pro * Independent			-0.0165	0.2220
GOP_Offline_Pro * Independent			-0.0386	0.0381
GOP_Online_Anti * Independent			-0.2462	-0.2670
GOP_Online_Pro * Independent			-0.1852	-0.0415
Dem_Offline_Pro * Republican			-0.2371	-0.2744
Dem_Online_Anti * Republican			-0.0531	-0.2091
Dem_Online_Pro * Republican			-0.2013	-0.1785
GOP_Offline_Pro * Republican			-0.0935	-0.1316
GOP_Online_Anti * Republican			-0.3455	-0.2652
GOP_Online_Pro * Republican			0.1188	0.1393

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 25: Preferences for Law Banning Offensive Online Speech

Comparison of Regression Models of Elite Cue Treatment Group Against Support of Law Banning Public Hate Speech

More positive numbers indicate support for law

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	3.540***	3.299***	3.483***	2.961***
treat				
Control	—	—	—	—
Dem_Offline_Pro	-0.0288	0.0211	-0.0331	-0.0368
Dem_Online_Anti	-0.0278	0.1027	-0.0456	0.1638
Dem_Online_Pro	-0.0066	0.0615	0.0843	0.0709
GOP_Offline_Pro	0.0896	0.1233	0.0940	0.1663
GOP_Online_Anti	0.0083	-0.0002	0.1674	0.1731
GOP_Online_Pro	-0.0107	0.0788	0.1490	0.1199
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-0.4052***	-0.1288*	-0.3775**	-0.0891
Republican	-0.6006***	-0.1999**	-0.4577***	0.1177
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		-0.7393***		-0.7335***
Gender				
Female		—		—
Male		-0.2535***		-0.2485***
Non-binary		-0.3100		-0.3189
Ethnicity				
White		—		—
Asian		0.0500		0.0578
Black		0.3979***		0.4053***
Latino		0.0096		0.0272
Multiracial		-0.0287		-0.0243
Native American		-0.0097		-0.0228
Age		0.0568***		0.0579***
Socials				
At least once a week but not every day		—		—
A few times a month		0.0306		0.0503
Every day		0.0985		0.1079
Less often or not at all		0.0815		0.0904
Community				
Suburb		—		—
City		-0.0620		-0.0654
Other		0.9209		0.9899
Rural		-0.0548		-0.0529
Town		0.1064		0.0981
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		-0.0152		-0.0170
Graduate degree		0.1076		0.0992
Prefer not to say		0.3434		0.3587
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			0.1160	0.2425
Dem_Online_Anti * Independent			0.0983	0.0373
Dem_Online_Pro * Independent			-0.0034	0.2028
GOP_Offline_Pro * Independent			0.0033	0.0180
GOP_Online_Anti * Independent			-0.0976	-0.1384
GOP_Online_Pro * Independent			-0.2920	-0.0001
Dem_Offline_Pro * Republican			-0.0767	0.0030
Dem_Online_Anti * Republican			-0.0247	-0.1838
Dem_Online_Pro * Republican			-0.2719	-0.1957
GOP_Offline_Pro * Republican			-0.0076	-0.1083
GOP_Online_Anti * Republican			-0.3929*	-0.3760
GOP_Online_Pro * Republican			-0.2218	-0.0902
Party_Affinity				0.0473

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 26: Preferences for Law Banning Public Hate Speech

Comparison of Regression Models of Elite Cue Treatment Group Against Perceived Relative Importance of Free Speech

Feeling Thermometer of 0 (Prefer Harm Protection) to 10 (Prefer Freedom of Speech)

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	6.096***	6.231***	6.079***	6.054***
treat				
Control	—	—	—	—
Dem_Offline_Pro	0.1478	0.0642	0.2491	0.2283
Dem_Online_Anti	-0.0689	-0.2660	0.1373	-0.1217
Dem_Online_Pro	-0.0979	-0.2328	-0.3933	-0.3369
GOP_Offline_Pro	-0.1623	-0.2056	-0.0558	0.0117
GOP_Online_Anti	-0.0819	-0.1092	-0.2399	-0.1277
GOP_Online_Pro	-0.0322	-0.1464	0.1397	0.3712
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	0.9585***	0.1136	0.8563**	0.2591
Republican	1.410***	-0.0461	1.539***	0.2556
Party_Affinity		-0.1177		-0.1079
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		1.537***		1.532***
Gender				
Female		—		—
Male		0.7827***		0.7846***
Non-binary		1.130*		1.137*
Ethnicity				
White		—		—
Asian		-0.2971		-0.2991
Black		-0.5108**		-0.5256**
Latino		-0.0225		-0.0537
Multiracial		0.4329*		0.4150
Native American		-1.299		-1.255
Age		0.0445		0.0440
Socials				
At least once a week but not every day		—		—
A few times a month		-0.4171		-0.3753
Every day		-0.1162		-0.1294
Less often or not at all		-0.0731		-0.1209
Community				
Suburb		—		—
City		0.3484*		0.3576**
Other		-0.4318		-0.3670
Rural		0.3086*		0.3246*
Town		0.0485		0.0668
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		0.2686		0.2641
Graduate degree		-0.2445		-0.2412
Prefer not to say		-0.7291		-0.8061
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			0.0372	-0.0025
Dem_Online_Anti * Independent			-0.2639	-0.0943
Dem_Online_Pro * Independent			0.5926	0.1042
GOP_Offline_Pro * Independent			-0.0596	-0.4309
GOP_Online_Anti * Independent			0.2524	0.2036
GOP_Online_Pro * Independent			0.1119	-0.6175
Dem_Offline_Pro * Republican			-0.3280	-0.4498
Dem_Online_Anti * Republican			-0.3961	-0.3280
Dem_Online_Pro * Republican			0.3970	0.2423
GOP_Offline_Pro * Republican			-0.2613	-0.2455
GOP_Online_Anti * Republican			0.2737	-0.0787
GOP_Online_Pro * Republican			-0.6157	-1.001*

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 27: Attitudes about the Value of Free Speech

Comparison of Regression Models of Elite Cue Treatment Group Against Favorability Towards Online Content Moderation

Feeling Thermometer of 0 (Coldest) to 100 (Warmest)

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**Beta**[†]	**Beta**[†]	**Beta**[†]	**Beta**[†]
(Intercept)	79.81***	67.53***	77.66***	64.00***
treat				
Control	—	—	—	—
Dem_Offline_Pro	-1.283	-2.437	1.309	2.493
Dem_Online_Anti	-1.795	-1.249	1.042	4.658
Dem_Online_Pro	-1.735	-2.344	0.7416	0.9526
GOP_Offline_Pro	-1.232	-3.563	0.2684	-0.9265
GOP_Online_Anti	-0.9218	-1.451	2.117	3.322
GOP_Online_Pro	-1.166	-0.4122	1.417	2.385
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-16.41***	-4.242	-15.13***	-1.435
Republican	-24.10***	-6.562	-18.82***	0.6524
Party_Affinity		1.631		1.606
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		-20.95***		-20.85***
Gender				
Female		—		—
Male		-6.613***		-6.625***
Non-binary		-5.733		-5.150
Ethnicity				
White		—		—
Asian		-0.9592		-0.8354
Black		6.322**		6.235**
Latino		2.235		2.416
Multiracial		-4.742*		-4.625*
Native American		-2.246		-2.578
Age		1.735***		1.749***
Socials				
At least once a week but not every day		—		—
A few times a month		0.0269		0.2791
Every day		-0.2115		0.0133
Less often or not at all		1.349		1.486
Community				
Suburb		—		—
City		2.555		2.535
Other		-3.680		-1.848
Rural		-1.440		-1.501
Town		1.636		1.610
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		-0.0634		-0.1580
Graduate degree		-0.8051		-0.8561
Prefer not to say		25.77		26.65
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			-2.002	-6.291
Dem_Online_Anti * Independent			-0.2533	-3.047
Dem_Online_Pro * Independent			1.439	-0.6132
GOP_Offline_Pro * Independent			-0.2179	-0.2747
GOP_Online_Anti * Independent			-0.9085	-4.438
GOP_Online_Pro * Independent			-6.692	-6.093
Dem_Offline_Pro * Republican			-6.007	-8.971
Dem_Online_Anti * Republican			-8.049	-14.10**
Dem_Online_Pro * Republican			-8.745*	-8.987
GOP_Offline_Pro * Republican			-4.208	-7.329
GOP_Online_Anti * Republican			-8.151	-9.947
GOP_Online_Pro * Republican			-1.705	-2.632

[†] *p<0.05; **p<0.01; ***p<0.001

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 28: Post-Treatment Online Content Moderation Preferences by post-treatment attention check Success

Comparison of Regression Models of Elite Cue Treatment Group Against Attention Check of Correctly Identifying Cue Content

1 = Correct Cue Content Identified; 0 = incorrect

	No Covariates	With Covariates	Interaction No Covariates	Interaction With Covariates
Variables	**log(OR)**[1,2]	**log(OR)**[1,2]	**log(OR)**[1,2]	**log(OR)**[1,2]
(Intercept)	0.4664***	0.4643	0.3635*	0.3933
treat				
Dem_Online_Pro	—	—	—	—
Dem_Offline_Pro	-0.0616	-0.2569	0.0652	-0.2293
Dem_Online_Anti	-1.122***	-1.380***	-0.9726***	-1.224***
GOP_Offline_Pro	-0.4189**	-0.4947**	-0.0991	-0.3204
GOP_Online_Anti	-0.8539***	-1.021***	-0.8004***	-0.9462***
GOP_Online_Pro	-0.1485	-0.2949	-0.1791	-0.4755
Grouped_Party_ID				
Democrat	—	—	—	—
Independent	-0.1668	-0.2587	0.2196	0.0598
Republican	-0.1488	-0.2385	-0.1737	-0.3942
Party_Affinity		0.0069		0.0092
Gov				
There are more things that government should be doing		—		—
Government is doing too many things better left to businesses and individuals		0.0859		0.0946
Gender				
Female		—		—
Male		0.1303		0.1462
Non-binary		-0.3960		-0.5165
Ethnicity				
White		—		—
Asian		0.0769		0.0853
Black		-0.5078**		-0.5087**
Latino		-0.5445		-0.5036
Multiracial		-0.0686		-0.0623
Native American		0.6217		0.6139
Age		-0.0088		-0.0061
Socials				
At least once a week but not every day		—		—
A few times a month		-0.1147		-0.0967
Every day		0.1037		0.0903
Less often or not at all		0.2687		0.2790
Community				
Suburb		—		—
City		0.0148		0.0256
Rural		0.0010		0.0079
Town		0.0404		0.0429
Grouped_Education				
High school or less		—		—
Bachelor's or Associates degree		0.0724		0.0909
Graduate degree		0.2051		0.1927
Prefer not to say		13.60		13.73
treat * Grouped_Party_ID				
Dem_Offline_Pro * Independent			-0.3511	-0.1778
Dem_Online_Anti * Independent			-0.7092*	-1.016*
GOP_Offline_Pro * Independent			-0.9272**	-0.7557
GOP_Online_Anti * Independent			-0.2842	-0.1794
GOP_Online_Pro * Independent			-0.1033	0.0816
Dem_Offline_Pro * Republican			-0.0726	0.0918
Dem_Online_Anti * Republican			0.1184	0.2810
GOP_Offline_Pro * Republican			-0.1606	0.1582
GOP_Online_Anti * Republican			0.0964	-0.0437
GOP_Online_Pro * Republican			0.1791	0.4677

[1] *p<0.05; **p<0.01; ***p<0.001

[2] OR = Odds Ratio

Original Prolific Study, conducted 2024 Sep 9-11, n = 3181

Figure 29: Post-treatment attention check

statements that have been made in the past. These hypothetical posts are modeled after posts that have appeared on social media, are not real, and do not represent the views of the researchers. You will be asked to imagine the generated statement you read is actually real and then answer some questions about the topic mentioned. Participation in this study is completely voluntary and the data will be fully anonymous; it is up to you to decide whether you want to participate or not. If you decide not to participate there will not be any negative consequences and no personal data will be stored. Please be aware that if you decide to participate, you may stop participating at any time. However, we can only reimburse those who complete the full questionnaire. The researchers will maintain the confidentiality of the research records or data, and all data will be stored at the University of Oxford by Oxford's Centre for Experimental Social Science. If you have a concern about any aspect of this study, please contact us. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, you may contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible (research@politics.ox.ac.uk).

Responsible members of the research team, data repositories, and funders may be given access to data for monitoring and/or audit of the study to ensure we are complying with guidelines, or as otherwise required by law. We plan to publish the results of this study in a scientific journal and access to the anonymous data might be given to journal editors and reviewers. Please note that you may only participate in this survey if you are 18 years of age or over and are an American citizen. If you have read the information above and agree to participate with the understanding that the data you submit will be processed accordingly, please check the box below to get started.

Thank you in advance for your participation!

Candidate XXXX

Dr. XXXX

Page Break

Statement of Consent By clicking Continue (->) below you are agreeing to the following statement:

I have read the information in this consent form. All my questions about the research have been answered to my satisfaction. If you do not wish to participate in the study, please exit the survey at this time.

Consent Do you provide your consent to participate in this study?

o Yes (1)

o No (2)

End of Block: Consent

Start of Block: Prolific ID

Q125 What is your Prolific ID?

Please note that this response should auto-fill with the correct ID

End of Block: Prolific ID

Start of Block: Pre-Treatment Demograpics

Pre_Online Please rate your feelings about social media companies' responsibility to moderate content on their online platforms using a scale from 0 to 100. A higher number indicates you strongly believe companies should moderate online content, while a lower number indicates you strongly believe they should not moderate online content. You can choose any number between 0 and 100.

Pre_Public Please rate your feelings about whether you believe government should prevent people from engaging in hate speech against certain groups in public using a scale from 0 to 100. A higher number means you strongly believe government should prevent public hate speech, while a lower number means you strongly believe they

should not. You can choose any number between 0 and 100.

Gov Which of these two statements comes closer to your own view?

o There are more things that government should be doing (1)

o Government is doing too many things better left to businesses and individuals (2)

Gender How do you describe yourself?

o Male (1)

o Female (2)

o Non-binary / third gender (3)

o Prefer to self-describe (4)

o Prefer not to say (5)

Ethnicity Choose one or more races/ethnicities that you consider yourself to be. Select one or more:

White (1)

Black or African American (2)

Asian (3)

American Indian or Alaska Native (4)

Latino/Hispanic (5)

Age How old are you?

o Under 18 (1)

o 18-24 years old (2)

o 25-34 years old (3)

o 35-44 years old (4)

o 45-54 years old (5)

o 55-64 years old (6)

o 65+ years old (7)

Party_ID Generally speaking, do you consider yourself as being a Republican, a Democrat or, an Independent?

o Strong Democrat (1)

o Democrat (2)

o Leaning Democrat (3)

o Independent (4)

o Leaning Republican (5)

o Republican (6)

o Strong Republican (7)

Socials Overall, how often would you say you visit social media platforms (Twitter, Facebook, etc.)?

o Every day (1)

o At least once a week but not every day (4)

o A few times a month (3)

o Less often or not at all (2)

Community What type of community do you live in?

o City (1)

o Suburb (2)

o Town (3)

o Rural (4)

o Other (5)

Education What is the highest level of education you have completed?

o Some high school or less (1)

o High school diploma or GED (2)

o Some college, but no degree (3)

o Associates or technical degree (4)

o Bachelor's degree (5)

o Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.) (6)

o Prefer not to say (7)

Timing_Pre Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

End of Block: Pre-Treatment Demograpics

Start of Block: Treatment Instructions

Treat Instructions Please read the following hypothetical statement carefully and then, imagining that the statement you read is actually a real statement, answer the following questions to the best of your abilities.

Start of Block: Post-treatment questionnaire

Q1 In your view, generally speaking, how should social media companies like Facebook and Twitter handle posts that violates their community guidelines?

o Leave them, do nothing (1)

o Place a warning label on these posts (2)

o Reduce how many people can see these posts (4)

o Permanently remove these posts (5)

o Suspend the accounts of the people who make these posts (6)

Q2 Please indicate who – if any – should be responsible for regulating content shared on these platforms (e.g., Facebook, Twitter, Instagram):

o There should be no regulation of content on social media (1)

o The social media companies should regulate content on social media (4)

o The national government should regulate content on social media (5)

o Both social media companies and the national government should regulate content on social media (6)

Q3 Would you favor or oppose a law that would make it illegal to say offensive or insulting things online to racial or religious groups?

o Strongly favor (2)

o Somewhat favor (3)

o Somewhat oppose (4)

o Strongly oppose (5)

o Don't know (6)

Post-Online On a scale from 0 to 100, how much responsibility do you think social media companies should have for regulating the content posted on their platforms? 0 indicates you firmly believe social media companies should not moderate content. 100 indicates you firmly believe social media companies should moderate content.

Post_Public On a scale from 0 to 100, how much do you agree with the idea that the government should restrict hate speech targeting specific groups in public spaces? 0 indicates you strongly believe the government should not restrict public hate speech. 100 indicates you strongly believe the government should restrict public hate speech.

Q4 Sometimes protestors organize protests at funerals specifically to cause emotional distress to the family of the deceased. Do you support the government passing a law to ban this type of protest?

o Strongly support (1)

o Somewhat support (2)

o Somewhat oppose (3)

o Strongly oppose (4)

o Don't know (5)

Q5 The U.S. Supreme Court has repeatedly ruled that hate speech – which attacks people based on their race, religion, gender identity or sexual orientation – is legally protected free speech by the First Amendment. Would you favor or oppose a change to law that would make it illegal to say hate speech in public to racial or religious groups?

o Strongly favor (1)

o Somewhat favor (4)

o Somewhat oppose (3)

o Strongly oppose (5)

o Don't know (6)

Q6 In general, how important is freedom of speech relative to the harm it might cause? Please give your response on a scale of 0 to 10, with 0 indicating you strongly prefer protection from harm and 10 indicating you strongly prefer freedom of speech.

Timing_Post Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

End of Block: Post-treatment questionnaire

Start of Block: Debrief

Debrief Thank you for taking the time to answer our questionnaire. As we said in the beginning, we fielded this study to understand citizens' political attitudes and attitudes about social media. Additionally, we implemented a brief experiment at the end of the survey to understand people's moderation preferences about social media content and public speech.

The goal of this study was to evaluate how attitudes about online content moderation and public speech change after exposure to statements from political leaders.

There is discussion among academics about what causes disagreement about contentious political issues. One particularly contentious issue is the role of authorities in moderating content shared online. This aims of this study is part of a larger effort to study people's attitudes about online content moderation and free speech in general.

In this study, some of you were randomly assigned to either a control group or one of six different treatment groups and then asked to answer the same survey questions. You were told that these statements, which were made to look like tweets, are hypothetical: that is none of them have really been tweeted and that you were asked to solely imagine as if they were real. A control group were not asked to read any hypothetical tweet and were instead asked to click "Continue" to directly answer the survey questions. A second group got a hypothetical, generated tweet in favor of social media content moderation that appeared to be from Democratic Senate Leadership. A third group got a hypothetical, generated tweet tweet in favor of social media content moderation that appeared to be from Republican Senate Leadership. A fourth group got a hypothetical, generated tweet against social media content moderation that appeared to be from Democratic Senate Leadership. A fifth group got a hypothetical, generated tweet against social media content moderation that appeared to

be from Republican Senate Leadership. A fifth group got a hypothetical, generated tweet in favor of countering harmful public speech that appeared to be from Democratic Senate Leadership. A sixth group got a hypothetical, generated tweet in favor of countering harmful public speech that appeared to be from Republican Senate Leadership.

All participants then answered the same survey questions. We asked you questions that would measure your preferences for the moderation of social media content and public speech to get an idea of what people think should be done about content moderation after a political elite group that you may or may not identify with makes a statement either in favor or against moderation.

In the next couple of years, we will be able to publish this work, and your responses are invaluable to us. We will be treating them anonymously (as you know, we are not collecting your names, or IP addresses), and it will be our pleasure to share our findings as soon as they are available.

Finally, we urge you not to discuss this study with anyone else who is currently participating or might participate at a future point in time since this might affect their responses.

Thank you!

Debrief Do you want to submit your answers?

o Yes (1)

o No (2)

End of Block: Debrief

Start of Block: Thank you

Q126 Thank you for taking part in this study. Please click the button below to be redirected back to Prolific and register your submission.

End of Block: Thank you

Start of Block: Attention Check

Q7 Who or which group is said to have made the statement you read earlier in this study?

o Senate Democrats (2)

o Senate Republicans (3)

o Forbes (4)

o Don't recall (5)

Q8 Recalling the statement you read on the previous page, about how much do you agree with the statement's message?

o Strongly agree (1)

o Somewhat agree (2)

o Neither agree nor disagree (4)

o Somewhat disagree (5)

o Strongly disagree (6)

o Don't recall (7)

Q9 What was the content of the statement you read earlier in this study?

o Pro online content moderation (1)

o Anti online content moderation (2)

o Pro moderation of public speech (3)

o Anti moderation of public speech (5)

o Don't recall (4)

Timing_Demographics Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

References

Katherine Adams, Jelani Cobb, Martin Baron, Russ Feingold, Lee C. Bollinger, Christina Paxson, Hillary Clinton, and Geoffrey R. Stone. Report of the Commission. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.002.0013. URL https://doi.org/10.1093/oso/9780197621080.002.0013**.**

Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.211. URL https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211**.**

Mauricio J. Alvarez and Markus Kemmelmeier. Free Speech as a Cultural Value in the United States. *Journal of Social and Political Psychology*, 5(2):707–735, 2017. ISSN 2195-3325. doi: 10.5964/jspp.v5i2.590. URL https://jspp.psychopen.eu/index.php/jspp/article/view/5031**. Number: 2.**

Emily A. Vogels Anderson, Andrew Perrin and Monica. Most Americans Think Social Media Sites Censor Political Viewpoints, August 2020. URL https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/**.**

Ruth E. Appel, Jennifer Pan, and Margaret E. Roberts. Partisan conflict over content moderation is more than disagreement about facts, July 2023. URL https://papers.ssrn.com/abstract=4331868**.**

Jack M. Balkin. To Reform Social Media, Reform Informational Capitalism. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0014. URL https://doi.org/10.1093/oso/9780197621080.003.0014**.**

Emily Bazelon. The Disinformation Dilemma. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0003. URL https://doi.org/10.1093/oso/9780197621080.003.0003**.**

Joe Biden. Statement by Vice President Kamala Harris, July 2024a. URL https://www.whitehouse.gov/briefing-room/statements-releases/2024/07/13/statement-by-vice-president-kamala-harris-2/**.**

Joe Biden. Statement from President Joe Biden, July 2024b. URL https://www.whitehouse.gov/briefing-room/statements-releases/2024/07/13/statement-from-president-joe-biden-6/**.**

Lee C. Bollinger and Geoffrey R. Stone. Opening Statement. In Lee C. Bollinger, Geoffrey R. Stone, Lee C. Bollinger, and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022a. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.002.0008. URL https://doi.org/10.1093/oso/9780197621080.002.0008**.**

Lee C. Bollinger and Geoffrey R. Stone. *Social media, freedom of speech, and the future of our democracy*. Oxford university press, New York, 2022b. ISBN 978-0-19-762109-7 978-0-19-762108-0.

Guy E. Carmi. Dignity Versus Liberty: The Two Western Cultures of Free Speech, August 2008. URL https://papers.ssrn.com/abstract=1246700**.**

Charlotte Cavaillé and Anja Neundorf. Elite Cues and Economic Policy Attitudes: The Mediating Role of Economic Hardship. *Political Behavior*, 45(4):1355–1376, December 2023. ISSN 1573-6687. doi: 10.1007/s11109-021-09768-w. URL https://doi.org/10.1007/s11109-021-09768-w.

Andrew J. Ceresney, Jeffrey P. Cunard, Courtney M. Dankworth, and David A. O'Neil. Regulating Harmful Speech on Social Media: The Current Legal Landscape and Policy Proposals. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.002.0009. URL https://doi.org/10.1093/oso/9780197621080.002.0009.

Erwin Chemerinsky and Alex Chemerinsky. The Golden Era of Free Speech. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0006. URL https://doi.org/10.1093/oso/9780197621080.003.0006.

U.S. Congress. 47 U.S. Code § 230 - Protection for private blocking and screening of offensive material, 1996. URL https://www.law.cornell.edu/uscode/text/47/230.

Renée DiResta. Algorithms, Affordances, and Agency. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0008. URL https://doi.org/10.1093/oso/9780197621080.003.0008.

Evelyn Douek. Content Moderation as Systems Thinking, December 2022. URL https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/.

James N. Druckman, Erik Peterson, and Rune Slothuus. How Elite Partisan Polarization Affects Public Opinion Formation. *American Politi-*

cal Science Review, 107(1):57–79, February 2013. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055412000500. URL https://www.cambridge.org/core/journals/american-political-science-review/article/abs/how-elite-partisan-polarization-affects-public-opinion-formation/6CB23BCCFBFBB4EA3879D91232CEEA59. Publisher: Cambridge University Press.

Randy Fisher, Stuart Lilie, Clarice Evans, Greg Hollon, Mary Sands, Dawn Depaul, Christine Brady, David Lindbom, Dawn Judd, Michelle Miller, and Tim Hultgren. Political Ideologies and Support for Censorship: Is It a Question of Whose Ox Is Being Gored? *Journal of Applied Social Psychology*, 29(8):1705–1731, 1999. ISSN 1559-1816. doi: 10.1111/j.1559-1816.1999.tb02047.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1999.tb02047.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.1999.tb02047.x.

James FitzGerald. Trump recounts shooting in marathon Republican convention speech, July 2024. URL https://www.bbc.com/news/articles/c4ngwmjedm3o. Section: US & Canada.

Mary Anne Franks. The Free Speech Industry. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0005. URL https://doi.org/10.1093/oso/9780197621080.003.0005.

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, 4(12):1285–1293, December 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-00994-6. URL https://www.nature.com/articles/s41562-020-00994-6. Number: 12 Publisher: Nature Publishing Group.

Dominique Geissler, Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, 12(1):1–20, December 2023. ISSN 2193-1127. doi: 10.1140/epjds/

s13688-023-00414-5. URL https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-023-00414-5**. Number: 1 Publisher: SpringerOpen.**

Robert Gorwa. Who Are the Stakeholders in Platform Governance? preprint, SocArXiv, September 2022. URL https://osf.io/ayx8h**.**

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, **7(1):2053951719897945, January 2020. ISSN 2053-9517. doi: 10.1177/2053951719897945. URL** https://doi.org/10.1177/2053951719897945**. Publisher: SAGE Publications Ltd.**

Jamal Greene. Free Speech on Public Platforms. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, **page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0010. URL** https://doi.org/10.1093/oso/9780197621080.003.0010**.**

James Grimmelmann. The Virtues of Moderation. *Cornell Law Faculty Publications*, **April 2015. URL** https://scholarship.law.cornell.edu/facpub/1486**.**

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science (New York, N.Y.)*, **363(6425):374–378, January 2019. ISSN 1095-9203. doi: 10.1126/science.aau2706.**

Jacob Poushter Gubbala, Moira Fagan and Sneha. Climate Change Remains Top Global Threat Across 19-Country Survey, August 2022. URL https://www.pewresearch.org/global/2022/08/31/climate-change-remains-top-global-threat-across-19-country-survey/**.**

Andrew M. Guess, Brendan Nyhan, and Jason Reifler. Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, **4(5):472–480, May 2020.**

ISSN 2397-3374. doi: 10.1038/s41562-020-0833-x. URL https://www.nature.com/articles/s41562-020-0833-x. Number: 5 Publisher: Nature Publishing Group.

Joseph Henrich. The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter. In *The Secret of Our Success*. Princeton University Press, October 2015. ISBN 978-1-4008-7329-6. doi: 10.1515/9781400873296. URL https://www.degruyter.com/document/doi/10.1515/9781400873296/html.

Ronald F. Inglehart. *Cultural Evolution*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-48931-7. doi: 10.1017/9781108613880. URL https://www.cambridge.org/core/books/cultural-evolution/34F637928AB1AA87B6409C28B4DFC9F5.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(1):129–146, May 2019. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-051117-073034. URL https://www.annualreviews.org/doi/10.1146/annurev-polisci-051117-073034.

Mike Johnson. Speaker Mike Johnson on Trump shooting: 'Everyone needs to turn the rhetoric down', 2024. URL https://www.nbcnews.com/politics/donald-trump/speaker-mike-johnson-trump-shooting-political-rhetoric-rcna161762.

Joshua D. Kertzer and Thomas Zeitzoff. A Bottom-Up Theory of Public Opinion about Foreign Policy. *American Journal of Political Science*, 61(3):543–558, July 2017. ISSN 0092-5853, 1540-5907. doi: 10.1111/ajps.12314. URL https://onlinelibrary.wiley.com/doi/10.1111/ajps.12314.

Amy Klobuchar. Profit Over People: How to Make Big Tech Work for Americans. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0018. URL https://doi.org/10.1093/oso/9780197621080.003.0018.

Spyros Kosmidis and Yannis Theocharis. Can Social Media Incivility Induce Enthusiasm?: Evidence from Survey Experiments. *Public Opinion Quarterly*, 84(S1): 284–308, April 2020. ISSN 0033-362X. doi: 10.1093/poq/nfaa014. URL https://doi.org/10.1093/poq/nfaa014.

Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan Herzog, Ullrich Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joseph Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam Berinsky, Cornelia Betsch, John Cook, Lisa Fazio, Michael Geers, Andrew Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson, Paula Szewach, Dr Sander van der Linden, and Sam Wineburg. Toolbox of Interventions Against Online Misinformation. January 2024. doi: 10.31234/osf.io/x8ejt. URL https://osf.io/x8ejt. Publisher: OSF.

Larry Kramer. A Deliberate Leap in the Opposite Direction: The Need to Rethink Free Speech. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0002. URL https://doi.org/10.1093/oso/9780197621080.003.0002.

Ronald J. Jr. Krotoszynski. Free Speech Paternalism and Free Speech Exceptionalism: Pervasive Distrust of Government and the Contemporary First Amendment Book Reviews and Responses. *Ohio State Law Journal*, 76(3):659–690, 2015. URL https://heinonline.org/HOL/P?h=hein.journals/ohslj76&i=665.

Genevieve Lakier. The Limits of Antidiscrimination Law in the Digital Public Sphere. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0011. URL https://doi.org/10.1093/oso/9780197621080.003.0011.

Lawrence Lessig. The First Amendment Does Not Protect Replicants. In Lee C.

Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0016. URL https://doi.org/10.1093/oso/9780197621080.003.0016.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of Hate Speech in Online Social Media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, pages 173–182, New York, NY, USA, June 2019. Association for Computing Machinery. ISBN 978-1-4503-6202-3. doi: 10.1145/3292522.3326034. URL https://doi.org/10.1145/3292522.3326034.

Paul Mena. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet*, 12(2):165–183, 2020. ISSN 1944-2866. doi: 10.1002/poi3.214. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.214. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.214.

Ryan C. Moore, Ross Dahlke, and Jeffrey T. Hancock. Exposure to untrustworthy websites in the 2020 US election. *Nature Human Behaviour*, 7(7):1096–1105, 2023. URL https://ideas.repec.org//a/nat/nathum/v7y2023i7d10.1038_s41562-023-01564-2.html. Publisher: Nature.

Mohsen Mosleh and David G. Rand. Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, 13(1):7144, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34769-6. URL https://www.nature.com/articles/s41467-022-34769-6. Number: 1 Publisher: Nature Publishing Group.

Kevin Munger. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, 39(3):629–649, September 2017. ISSN 1573-6687. doi: 10.1007/s11109-016-9373-5. URL https://doi.org/10.1007/s11109-016-9373-5.

Diana C. Mutz. Freedom of Speech in the Post-Floyd Era: Public Support for Political Tolerance. *The ANNALS of the American Academy of Political and Social Science*, 708(1):184–205, July 2023. ISSN 0002-7162. doi: 10.1177/00027162241231129. URL https://doi.org/10.1177/00027162241231129. Publisher: SAGE Publications Inc.

Karsten Müller and Carlo Schwarz. Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, 19(4):2131–2167, August 2021. ISSN 1542-4766. doi: 10.1093/jeea/jvaa045. URL https://doi.org/10.1093/jeea/jvaa045.

Stephen P. Nicholson. Dominating Cues and the Limits of Elite Influence. *The Journal of Politics*, 73(4):1165–1177, October 2011. ISSN 0022-3816, 1468-2508. doi: 10.1017/S002238161100082X. URL https://www.journals.uchicago.edu/doi/10.1017/S002238161100082X.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Canto Classics. Cambridge University Press, Cambridge, 2015. ISBN 978-1-107-56978-2. doi: 10.1017/CBO9781316423936. URL https://www.cambridge.org/core/books/governing-the-commons/A8BB63BC4A1433A50A3FB92EDBBB97D5.

Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, July 2019. ISSN 1873-7838. doi: 10.1016/j.cognition.2018.06.011.

Nathaniel Persily. Platform Power, Online Speech, and the Search for New Constitutional Categories. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0012. URL https://doi.org/10.1093/oso/9780197621080.003.0012.

Eric A. Posner. Symbols, Signals, and Social Norms in Politics and the Law. *The Journal of Legal Studies*, 27(S2):765–797, June 1998. ISSN 0047-2530. doi: 10.1086/468042. URL https://www.journals.uchicago.edu/doi/abs/10.1086/468042. Publisher: The University of Chicago Press.

Franziska Pradel, Jan Zilinsky, Spyros Kosmidis, and Yannis Theocharis. Do Users Ever Draw a Line? Offensiveness and Content Moderation Preferences on Social Media. December 2023. doi: 10.31219/osf.io/y4xft. URL https://osf.io/y4xft. Publisher: OSF.

Franziska Pradel, Jan Zilinsky, Spyros Kosmidis, and Yannis Theocharis. Toxic Speech and Limited Demand for Content Moderation on Social Media. *American Political Science Review*, **pages 1–18, January 2024. ISSN 0003-0554, 1537-5943. doi: 10.1017/S000305542300134X. URL** https://www.cambridge.org/core/journals/american-political-science-review/article/toxic-speech-and-limited-demand-for-content-moderation-on-social-media/405333D7072585903E81BEF1729378F8**.**

Martin Redish. Value of Free Speech. *University of Pennsylvania Law Review*, **130 (3):591, January 1982. URL** https://scholarship.law.upenn.edu/penn_law_review/vol130/iss3/2**.**

David Richards. Free Speech and Obscenity Law: Toward a Moral Theory of the First Amendment. *University of Pennsylvania Law Review*, **123(1):45, November 1974. URL** https://scholarship.law.upenn.edu/penn_law_review/vol123/iss1/2**.**

Jon Roozenbeek, Rakoen Maertens, Stefan M. Herzog, Michael Geers, Ralf Kurvers, Mubashir Sultan, and Sander van der Linden. Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, **17(3):547–573, 2022. URL** https://econpapers.repec.org/article/cupjudgdm/v_3a17_3ay_3a2022_3ai_3a3_3ap_3a547-573_5f3.htm**. Publisher: Cambridge University Press.**

M. Rungtusanatham, Cynthia Wallin, and Stephanie Eckerd. The Vignette in a Scenario-Based Role-Playing Experiment. *Journal of Supply Chain Management*, **47(3):9–16, 2011. ISSN 1745-493X. doi: 10.1111/j.1745-493X.2011.03232.x. URL** https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-493X.2011.03232.x**. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-493X.2011.03232.x.**

Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. A Web of Hate: Tackling Hateful Speech in Online Social Spaces, September 2017. URL http://arxiv.org/abs/1709.10159**. arXiv:1709.10159 [cs].**

Steve Scalise. Scalise Statement on Shooting at Trump Rally | Congressman Steve Scalise, July 2024. URL http://scalise.house.gov/press-release/Scalise-Statement-on-Shooting-at-Trump-Rally**.**

Michael Scherer. Trump allies immediately blame Biden, Democrats for their rhetoric. *Washington Post*, July 2024. ISSN 0190-8286. URL https://www.washingtonpost.com/politics/2024/07/13/trump-shooting-blame-biden-democrats/**.**

ALEXANDRA SIEGEL and Vivienne Badaan. #No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online. *American Political Science Review*, 114:837–855, June 2020. doi: 10.1017/S0003055420000283.

Richard Wike and Katie Simmons. Global Support for Principle of Free Expression, but Opposition to Some Forms of Speech, November 2015. URL https://www.pewresearch.org/global/2015/11/18/global-support-for-principle-of-free-expression-but-opposition-to-some-forms-of-speec

Rasmus Skytte. Dimensions of Elite Partisan Polarization: Disentangling the Effects of Incivility and Issue Polarization. *British Journal of Political Science*, 51(4):1457–1475, 2021. URL https://ideas.repec.org//a/cup/bjposi/v51y2021i4p1457-1475_6.html**. Publisher: Cambridge University Press.**

Paul M. Sniderman, Richard A. Brody, and Philip E. Tetlock, editors. *Reasoning and choice: explorations in political psychology*. Cambridge Univ. Pr, Cambridge, paperback ed., repr., digital print edition, 2003. ISBN 978-0-521-40770-0 978-0-521-40255-2.

Brittany C. Solomon, Matthew E. K. Hall, Abigail Hemmen, and James N. Druckman. Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it. *Proceedings of the National Academy of Sciences*, 121(39):e2402428121, September 2024. doi: 10.1073/pnas.2402428121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2402428121**. Publisher: Proceedings of the National Academy of Sciences.**

Søren Staghøj. Who Cares About Free Speech? 2021.

Kate Starbird. Strategy and Structure: Understanding Online Disinformation and How Commitments to "Free Speech" Complicate Mitigation Approaches. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0013. URL https://doi.org/10.1093/oso/9780197621080.003.0013.

David A. Strauss. Social Media and First Amendment Fault Lines. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0001. URL https://doi.org/10.1093/oso/9780197621080.003.0001.

Cass R. Sunstein. A Framework for Regulating Falsehoods. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0004. URL https://doi.org/10.1093/oso/9780197621080.003.0004.

Karolina Sylwester and Matthew Purver. Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *PLOS ONE*, 10(9):e0137422, September 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0137422. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137422. Publisher: Public Library of Science.

Henri Tajfel and John C. Turner. *The Social Identity Theory of Intergroup Behavior*. Political psychology: Key readings. Psychology Press, New York, NY, US, 2004. ISBN 978-1-84169-069-8 978-1-84169-070-4. doi: 10.4324/9780203505984-16. Pages: 293.

Kai M. Thaler. Mixed Methods Research in the Study of Political and Social Violence and Conflict. *Journal of Mixed Methods Research*, 11(1):59–76, January 2017. ISSN 1558-6898. doi: 10.1177/1558689815585196. URL https://doi.org/10.1177/1558689815585196. Publisher: SAGE Publications.

Richard H. Thaler and Cass R. Sunstein. *Nudge: the final edition*. Penguin Books, an imprint of Penguin Random House LLC, New York, updated edition edition, 2021. ISBN 978-0-14-313700-9.

Melania Trump. https://t.co/IGIWzL6SMJ, July 2024. URL https://x.com/MELANIATRUMP/status/1812492817068945437.

Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131, September 1974. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.185.4157.1124. URL https://www.science.org/doi/10.1126/science.185.4157.1124.

Eugene Volokh. Free Speech Rules, Free Speech Culture, and Legal Education. *Hofstra Law Review*, 51(3), June 2023. ISSN 00914029. URL https://scholarlycommons.law.hofstra.edu/hlr/vol51/iss3/5.

Thomas G. West. Free Speech in the American Founding and in Modern Liberalism. *Social Philosophy and Policy*, 21(2):310–384, 2004. doi: 10.1017/s0265052504212110. Publisher: Cambridge University Press.

Philippe Weyrich, Anna Scolobig, Florian Walther, and Anthony Patt. Do intentions indicate actual behaviour? A comparison between scenario-based experiments and real-time observations of warning response. *Journal of Contingencies and Crisis Management*, 28(3):240–250, 2020. ISSN 1468-5973. doi: 10.1111/1468-5973.12318. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-5973.12318. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-5973.12318.

Sheldon Whitehouse. Section 230 Reforms. In Lee C. Bollinger and Geoffrey R. Stone, editors, *Social Media, Freedom of Speech, and the Future of our Democracy*, page 0. Oxford University Press, August 2022. ISBN 978-0-19-762108-0. doi: 10.1093/oso/9780197621080.003.0007. URL https://doi.org/10.1093/oso/9780197621080.003.0007.

R. George Wright. Freedom of Speech as a Cultural Holdover, June 2019. URL https://papers.ssrn.com/abstract=3399624**.**

John R. Zaller. ***The nature and origins of mass opinion*****. Cambridge Univ. Press, Cambridge, 13. printing edition, 2006. ISBN 978-0-521-40786-1 978-0-521-40449-5.**