Resolving content moderation dilemmas between free speech and harmful misinformation

Kozyreva, A.; Herzog, S.M.; Lewandowsky, S.; Hertwig, R.; Lorenz-Spreen, P.; Leiser, M.R.; Reifler, J.

Free speech vs. harmful misinformation: Moral dilemmas in online content moderation

Anastasia Kozyreva[1*], Stefan M. Herzog[1], Stephan Lewandowsky[2,3], Ralph Hertwig[1], Philipp Lorenz-Spreen[1], Mark Leiser[4], and Jason Reifler[5]

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany, 14195

[2]School of Psychological Science, University of Bristol, Bristol, United Kingdom, BS8 1QU

[3]School of Psychological Sciences, University of Western Australia, Perth, Australia, 6009

[4]Center for Law and Digital Technologies, Leiden University, Leiden, Netherlands, 2311 ES

[5]Department of Politics, University of Exeter, Exeter, United Kingdom, EX4 4PY

[*]Corresponding author: Anastasia Kozyreva, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany; kozyreva@mpib-berlin.mpg.de

Abstract

When moderating content online, two key values may come into conflict: protecting freedom of expression and preventing harm. Robust rules based in part on how citizens think about these moral dilemmas are necessary to deal with the unprecedented scale and urgency of this conflict in a principled way. Yet little is known about people's judgments and preferences around content moderation. We examined such moral dilemmas in a conjoint survey experiment where respondents ($N = 2,564$) indicated whether they would remove problematic social media posts on election denial, anti-vaccination, Holocaust denial, and climate change denial and whether they would take punitive action against the accounts. Respondents were shown key information about the user and their post, as well as the consequences of the misinformation. The majority preferred quashing harmful misinformation over protecting free speech. Respondents were more likely to remove posts and suspend accounts if the consequences were severe and if it was a repeated offence. Features related to the account itself (the person behind the account, their partisanship, and number of followers) had little to no effect on respondents' decisions. Content moderation of harmful misinformation was a partisan issue: Across all four scenarios, Republicans were consistently less willing than Democrats or Independents to delete posts or penalize the accounts that posted them. Our results can inform the design of transparent rules of content moderation for human and algorithmic moderators.

Keywords: moral dilemma, freedom of expression, misinformation, disinformation, online speech, content moderation, harmful content, conjoint experiment

Free speech vs. harmful misinformation: Moral dilemmas in online content moderation

> We have a right to speak freely. We also have a right to life. When malicious disinformation—claims that are known to be both false and dangerous—can spread without restraint, these two values collide head-on.
>
> — George Monbiot, 2021

> The reality is we make a lot of decisions that affect people's ability to speak. [. . .] Frankly, I don't think we should be making so many important decisions about speech on our own either.
>
> — Mark Zuckerberg, 2019

Every day, human moderators and automated tools make countless decisions about what social media posts can be shown to users and what gets taken down, as well as how to discipline offending accounts. The ability to make these content moderation decisions at scale, thereby controlling online speech, is unprecedented in human history. Legal requirements make some content removal decisions easy for platforms and content moderators (e.g., selling illegal drugs or promoting terrorism). But what about when content is not explicitly illegal but rather "legal but harmful" or "lawful but awful"? Harmful misinformation—inaccurate claims that can cause harm—is one case in point. False and misleading information is considered harmful when it undermines people's ability to make informed choices and when it leads to adverse consequences such as threats to public health or to the legitimacy of an election (European Commission, 2020).

The scale and urgency of the problems around content moderation became particularly apparent when Donald Trump and political allies spread false information attacking the legitimacy of the 2020 presidential election, culminating in a violent attack on the U.S. Capitol. Subsequently, most major social media platforms suspended Trump's accounts (Clegg, 2021; Twitter, 2021; YouTube Insider, 2021). After a sustained period of prioritizing free speech and avoiding the role of "arbiters of truth" (Zuckerberg, 2016, 2019), social media platforms appear to be rethinking their approach to governing online speech (Douek, 2021). During the COVID-19 pandemic, most global social media platforms took an unusually interventionist approach to false information and vowed to remove or limit COVID-19 misinformation and conspiracies (Google, n.d.-a; Instagram, n.d.; Rosen, 2020; Twitter, n.d.-a). In 2020, Meta overturned its policy of allowing Holocaust denial, and removed some white supremacists groups from Facebook (Bickert, 2020); Twitter implemented a similar policy soon after (Twitter, n.d.-b). In October 2021, Google announced a policy forbidding advertising content on its platforms that "mak[es] claims that are demonstrably false and could significantly undermine participation or trust in an electoral or democratic process" or that "contradict[s] authoritative, scientific consensus on climate change" (Google, n.d.-b). And most recently, Pinterest introduced a new policy against false or misleading climate change information across both content and ads (Pinterest, n.d.). (For an overview of major platforms' moderation policies related to misinformation, see Appendix Table D1.)

At the core of these decisions is a moral dilemma: Should freedom of expression be upheld even at the expense of allowing dangerous misinformation to spread, or should misinformation be removed or penalized, thereby limiting free speech? When choosing between action (e.g., removing a post) and inaction (e.g., allowing a post to remain online), decision makers face a choice between two values (e.g., public health vs. freedom of expression) that, while not in themselves mutually exclusive, cannot be honored simultaneously. These cases are *moral dilemmas*: "situations where an agent morally ought to adopt each of two alternatives but cannot adopt both" (Sinott-Armstrong, 1988, p.5).

Although moral dilemmas have long been used in empirical studies of ethics and moral decision making, moral dilemmas in online content moderation are relatively new. So far, little is known about how people approach such latter dilemmas. Here we begin to bridge this gap by studying public content moderation preferences and investigating what attributes of content moderation dilemmas impact people's decisions the most.

Resolving content moderation dilemmas is difficult. Mitigating harms from misinformation by removing content and deplatforming accounts (especially at scale) might challenge the fundamental human right to "receive and impart information and ideas through any media and regardless of frontiers" (United Nations, 1948, art. 19). Moreover, there are good reasons within existing legal systems to protect even false speech (see Sunstein, 2021). People with the power to regulate speech based on its accuracy may succumb to the temptation to suppress opposition voices (e.g., authoritarian rulers often censor dissent by determining what is "true"). Censoring falsehoods might also prevent people from freely sharing their opinions, thereby deterring (e.g., due to fear of punishment) even legally protected speech (see Schauer, 1978). Indeed, a core tenet of the marketplace of ideas is that it can appropriately discard false and inaccurate claims: "The best test of truth is the power of an idea to get itself accepted in the competition of the market" (*Abrams v. United States*, 1919).

Do digital and social media, where harmful misinformation can quickly proliferate and where information flow is algorithmically moderated, belie this confidence in the marketplace of ideas? As Sunstein (2021) points out, "far from being the best test of truth, the marketplace ensures that many people accept falsehoods" (p. 49). For instance, when a guest on Joe Rogan's popular podcast shared discredited claims about COVID-19 vaccines, he spread potentially fatal misinformation to millions of listeners (Yang, 2022). Here two important points must be distinguished: First, while some types of misinformation may be relatively benign, others are harmful to people and the planet. For example, relative to factual information, exposure to misinformation can reduce people's intention to get vaccinated against COVID-19 by more than 6 percentage points (Loomba et al., 2021). This fact potentially invokes the principle of harm (Mill, 1859/2011; van Mill, 2021), which can justify limiting freedom of expression to prevent direct and imminent harms to others. Second, sharing one's private opinions, however unfounded, with a friend is substantially different from deliberately sharing potentially harmful falsehoods with virtually unlimited audiences. One may therefore argue that freedom of speech does not entail freedom of reach (Cohen, 2019), and that the right to express one's opinions is subject to limitations when the speech in question is amplified online.

Freedom of expression is an important right, and restrictions on false speech in liberal democracies are few and far between. State censorship is a trademark of authoritarianism: The Chinese government's censorship of internet content is a case in point (King et al., 2014), as is the introduction of "fake news" laws during the pandemic as a way for authoritarian states to justify repressive policies that stifle the opposition and further infringe on freedom of the press (The Economist, 2021; Wiseman, 2020; Yadav et al., 2021; for an overview of misinformation actions worldwide see Funke and Flamini, n.d.). Furthermore, in March 2022, the Russian parliament approved jail terms of up to 15 years for sharing "fake" (i.e., contradicting the official government position) information about the war against Ukraine, which led many foreign and local journalists and news organizations to limit their coverage of the invasion or withdraw from the country entirely.

Unlike in authoritarian or autocratic countries, in liberal democracies it is the platforms themselves that are the primary regulators of online speech. This responsibility raises the problem of rule-making powers being concentrated in the hands of a few unelected individuals at profit-driven companies. Furthermore, platforms increasingly rely on automated content moderation; for instance, the majority of hate speech on Facebook is removed by machine-learning algorithms (The Economist, 2020). Algorithmic content moderation at scale (Gorwa et al., 2020) poses additional challenges to an already complicated issue, including the inevitable occurrence of false positives, when acceptable content is removed, and false negatives, when posts violate platform policies but escape deletion. Algorithms operate on the basis of explicit and implicit rules (e.g., should they remove false information about climate change or only about COVID-19?). Content moderation—either purely algorithmic or with humans in the loop—inevitably requires a systemic balancing of individual speech rights against other societal interests and values

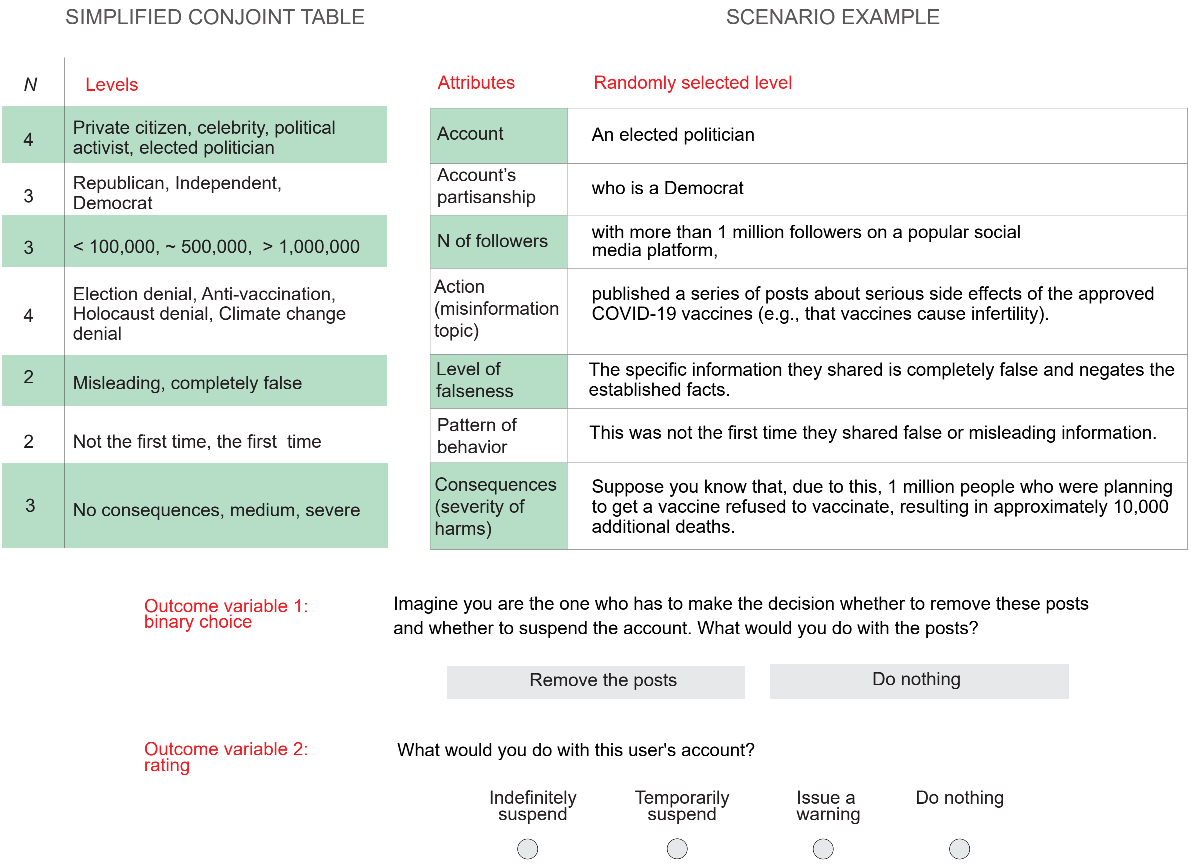
(Douek, 2021).

Experiments featuring moral dilemmas are an established approach to understanding people's moral intuitions around algorithmic decision making (Awad et al., 2018; Bonnefon et al., 2016) and computational ethics (Awad et al., 2022). Scenarios involving moral dilemmas (e.g., the trolley problem) are used widely in moral psychology to assess people's moral intuitions and reasoning (e.g., Greene et al., 2001). Classical moral dilemmas include scenarios involving choices between two obligations arising from the same moral requirement or from two different moral requirements. Most studies focus on moral dilemmas of the sacrificial type: presenting a choice within one moral requirement (e.g., saving lives) with asymmetrical outcomes (e.g., to save five lives by sacrificing one; see Foot, 1967; Thomson, 1985). In the case of content moderation decisions, however, we are dealing with dilemmas between two different values or moral requirements (e.g., protecting freedom of expression vs. mitigating potential threats to public health) that are incommensurate and whose adverse outcomes are difficult to measure or quantify.

We constructed four types of hypothetical scenarios arising from four contemporary topics that are hotbeds of misinformation: politics ("election denial" scenario), health ("anti-vaccination" scenario), history ("Holocaust denial" scenario), and the environment ("climate change denial" scenario). In designing these scenarios, we relied on the current content moderation policies of major social media platforms and selected topics where active polices have already been implemented (Appendix Table D1).

We used a single-profile conjoint survey experiment to explore what factors influence people's willingness to remove false and misleading content on social media and to penalize accounts that spread it. A conjoint design is particularly suitable for such a multilevel problem, where a variety of factors can impact decision-making (Bansak et al., 2021; Hainmueller, Hopkins, et al., 2014). Factors we focused on are: characteristics of the account (the person behind it, their partisanship, and the number of followers they have); characteristics of the shared content (the misinformation topic and whether it was completely false or only misleading); whether this was a repeated offence (i.e., a proxy for intent); and the consequences of sharing the information. All these factors were represented as attributes with distinct levels (Figure 1). This design yielded 1,728 possible unique cases.

Figure 1: Conjoint Scenario Design



SIMPLIFIED CONJOINT TABLE

N	Levels
4	Private citizen, celebrity, political activist, elected politician
3	Republican, Independent, Democrat
3	< 100,000, ~ 500,000, > 1,000,000
4	Election denial, Anti-vaccination, Holocaust denial, Climate change denial
2	Misleading, completely false
2	Not the first time, the first time
3	No consequences, medium, severe

SCENARIO EXAMPLE

Attributes	Randomly selected level
Account	An elected politician
Account's partisanship	who is a Democrat
N of followers	with more than 1 million followers on a popular social media platform,
Action (misinformation topic)	published a series of posts about serious side effects of the approved COVID-19 vaccines (e.g., that vaccines cause infertility).
Level of falseness	The specific information they shared is completely false and negates the established facts.
Pattern of behavior	This was not the first time they shared false or misleading information.
Consequences (severity of harms)	Suppose you know that, due to this, 1 million people who were planning to get a vaccine refused to vaccinate, resulting in approximately 10,000 additional deaths.

Outcome variable 1: binary choice

Imagine you are the one who has to make the decision whether to remove these posts and whether to suspend the account. What would you do with the posts?

Remove the posts | Do nothing

Outcome variable 2: rating

What would you do with this user's account?

Indefinitely suspend	Temporarily suspend	Issue a warning	Do nothing
○	○	○	○

Note. Complete phrasings of all attribute levels are listed in Table 1.

In the conjoint task, each respondent (N = 2,564) faced four random variations of each of the four scenario types (see Figure 1 for an example). These four scenario types represent four misinformation topics, with consequences adjusted for each topic. Each respondent evaluated 16 cases (40,845 evaluations in total). For each case, they were asked to make two choices: whether to remove the posts mentioned in the scenario and whether to suspend the account that posted them. We recruited 2,564 U.S. respondents via the Ipsos panel provider between October 18th and December 3rd 2021. The sample was quota-matched to the U.S. general population. The full experimental design and sample information are described in the Materials and Methods section.

Results

Restricting Misinformation: Decisions to Remove Posts and Penalize Accounts

For the majority of cases, across all four misinformation topics, respondents chose to remove posts featuring false or misleading information (Figure 2A). Climate change denial was removed the least (58%), whereas Holocaust denial was removed the most (71%), closely followed by election denial (69%) and anti-vaccination content (66%). In deciding the fate of an offending account (Figure 2C), respondents preferred to issue a warning (between 31% and 37%). However, the total amount of choices to temporarily or indefinitely suspend an account constituted about half of responses in the Holocaust denial (51%) and election denial (49%) scenarios, followed by the anti-vaccination (44%) and climate change denial scenarios (35%).

Figure 2B and D shows a clear difference in the proportion of choices to remove content or suspend accounts between Democrats and Republicans, with Independents in between. Only a small minority of Democrats chose to leave misinformation in place or opted to take action against the account spreading it. Republicans were almost evenly split in their decisions to remove the posts in three of the four scenarios; in the climate change denial scenario, a majority of Republican respondents preferred to do nothing.

Conjoint Analyses: What Influences Content Moderation Decisions?

To analyze respondents' content moderation preferences related to different conjoint factors, we computed average marginal component effects (AMCEs) for both outcome variables: the binary choice to remove the posts and the rating of how to handle the accounts. Figure 3 shows pooled results across all scenarios (i.e., the four scenario types are treated as the levels of the "misinformation topic" attribute; see Table 1 in Methods).

Three attributes had the largest effects on people's content removal decisions:

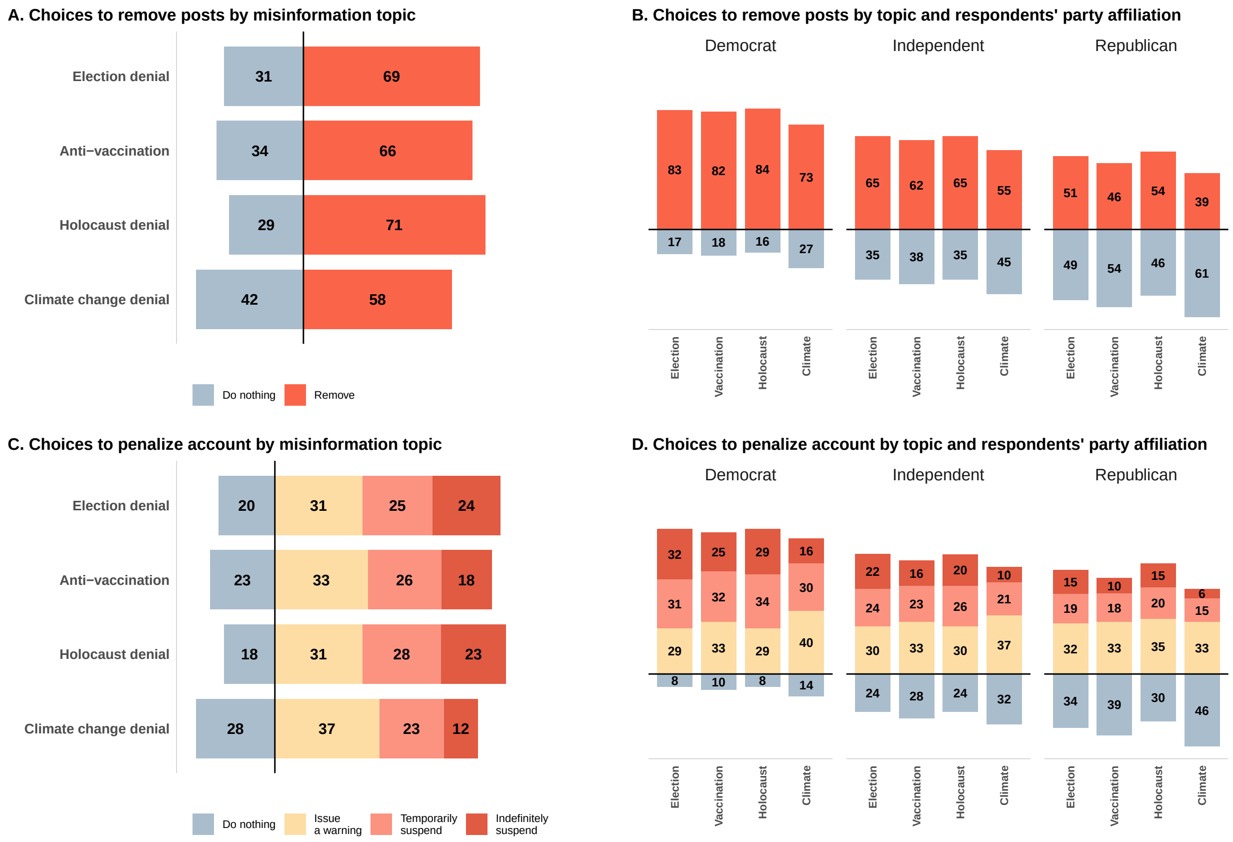
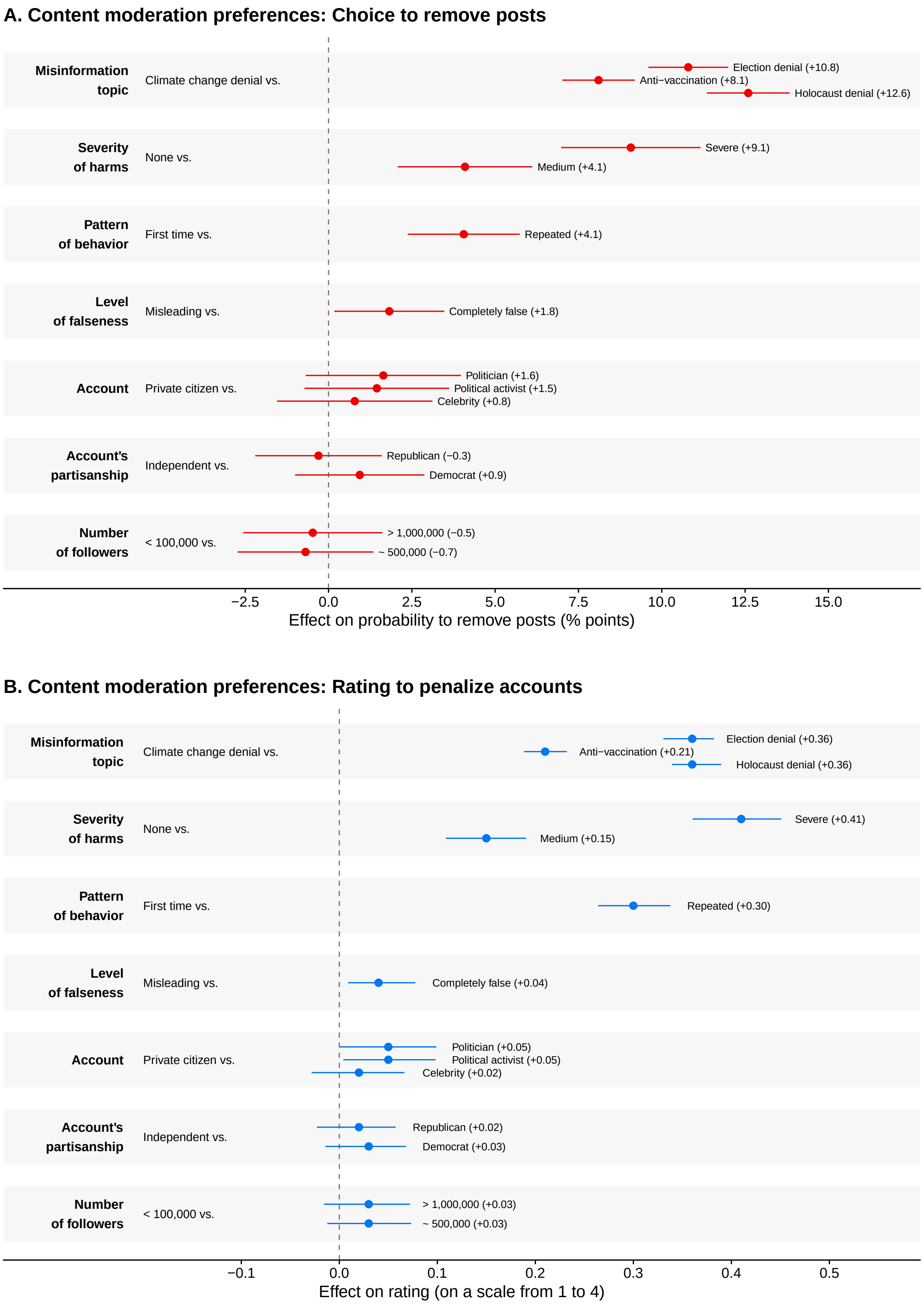


Figure 2: *Proportion of choices to remove posts and to suspend accounts.* All numeric values represent percentages. Panel A: Choices to remove posts or do nothing by misinformation topic (all cases). Panel B: Choices to remove posts or do nothing, by topic and respondents' party affiliation. Panel C: Choices to penalize account by misinformation topic (all cases). Panel D: Choices to penalize account by topic and respondents' party affiliation. $N = 40,845$ evaluated in total. (Cases evaluated by Democrats $n = 19,338$; by Independents $n = 8,229$; by Republicans $n = 13,278$.)

misinformation topic, severity of harm, and pattern of behavior. The misinformation topic consistently produces the largest effect. As Figure 3 demonstrates, changing the misinformation topic from climate change denial to Holocaust denial increased the probability to remove the posts by 13 percentage points and increased the rating to penalize accounts in the magnitude of 0.4 points on the 4-point scale (or 13% of the scale).

The second-strongest effect relates to the severity of harm: The more harmful the consequences of sharing misinformation (e.g., lives lost), the more likely respondents were to remove the posts. For instance, changing the severity of consequences from none to severe across scenarios increased the probability of choosing to remove the posts by 9



A. Content moderation preferences: Choice to remove posts

B. Content moderation preferences: Rating to penalize accounts

Figure 3: *Preferences for content moderation.* The figure reports average marginal component effects (AMCEs) plotted with 95% confidence intervals. In each row, effect sizes show an impact of each attribute level (on the right) relative to the reference attribute level (on the left), aggregated over all other attributes. Panel A: AMCEs are converted to percentage points and represent effects on probability to remove the posts. Panel B: AMCEs represent effects on rating to penalize the account. For marginal means, see Appendix Figure A1. For all AMCE and marginal means estimates, see Appendix Tables A3, A4, A5, A6.

percentage points (Figure 3A) and increased the penalization rating by almost half a rating step on the 4-point scale (Figure 3B).

A third important factor was the pattern of behavior. Changing this attribute from first offense to repeated offense increased the probability to remove the posts by 4 percentage points and increased the rating to penalize account in the magnitude of 0.3 points on the 4-point scale.

In sum, for decisions on both posts and accounts, severity of outcomes, repeated offence, and misinformation topic had the strongest impact on decisions to remove misinformation, while attributes related to an account's features, including the person behind it, their partisanship, and the number of followers, had little impact on participants' decisions. Whether the information was misleading or completely false was also relatively unimportant.

Subgroup Analyses: Effects of Partisanship and Attitudes Toward Free Speech

In order to assess how these content moderation preferences differed depending on respondents' attitudes, we conducted subgroup analyses for two main characteristics of interest: respondents' political partisanship and their attitude toward freedom of expression. Figure 4 shows marginal means and AMCEs for the choice to remove the posts for three subgroups: Republicans, Independents, and Democrats (see Table A1 for their distribution in the sample). Figure 5 shows marginal means and AMCEs for the choice to remove the posts for two subgroups: pro-freedom of expression and pro-mitigating harmful misinformation. These subgroups were formed based on responses to our pretreatment question: "If you absolutely have to choose between protecting freedom of expression and preventing disinformation from spreading, which is more important to you?" (Figure B1).

The AMCEs in Figure 4B show how different attribute levels affected the probability to remove the posts by respondents' party affiliation. All three groups showed similar patterns, with three exceptions. First, Republicans were, on average, more influenced than

Democrats or Independents by severity of outcomes and misinformation topic. Second, contrary to our expectations, there was no clear indication of a partisanship effect; that is, participants were not inclined to penalize an account that was at odds with their political leaning. Third, a large number of followers (i.e., $> 1{,}000{,}000$ relative to the reference level of $< 100{,}000$) had opposite effects on Republicans and Democrats (and no effects on Independents): For Democrats, a bigger reach increased the probability to remove the posts by 3 percentage points, whereas for Republicans it decreased the probability by 3 percentage points. Similar effects for this attribute appeared in the subgroup analyses for attitudes toward freedom of expression (see Figure 5): Respondents who valued freedom of expression over mitigating harmful misinformation were less likely to remove posts by accounts with many followers, whereas respondents who indicated that preventing misinformation was more important than protecting free speech were more likely to penalize accounts with many followers. More Republicans were pro-freedom of expression and more Democrats were pro-mitigating harmful misinformation (Appendix Figure B1).

Marginal means in Figure 5A show that participants made decisions that were consistent with their attitudes. On average, those who were pro-freedom of expression were equally or less likely to remove posts than they were to do nothing, whereas those who were pro-mitigating misinformation were much more likely to remove the posts than they were to do nothing. Marginal means in Figure 4A show that three partisan subgroups have different content moderation preferences. Republicans were least likely to remove posts for all attribute levels (but close to equal likelihood to remove or do nothing), while Democrats and Independents were more likely to remove the posts than to do nothing. The only attributes' levels that made Republicans more likely to remove the posts rather than leave them up were Holocaust denial content and posts with severe consequences.

Towards the end of the survey we assessed respondents' beliefs regarding a variety of claims relevant to our scenarios in order to better understand the role of accuracy of respondents' existing knowledge. Republicans were more likely than Democrats and

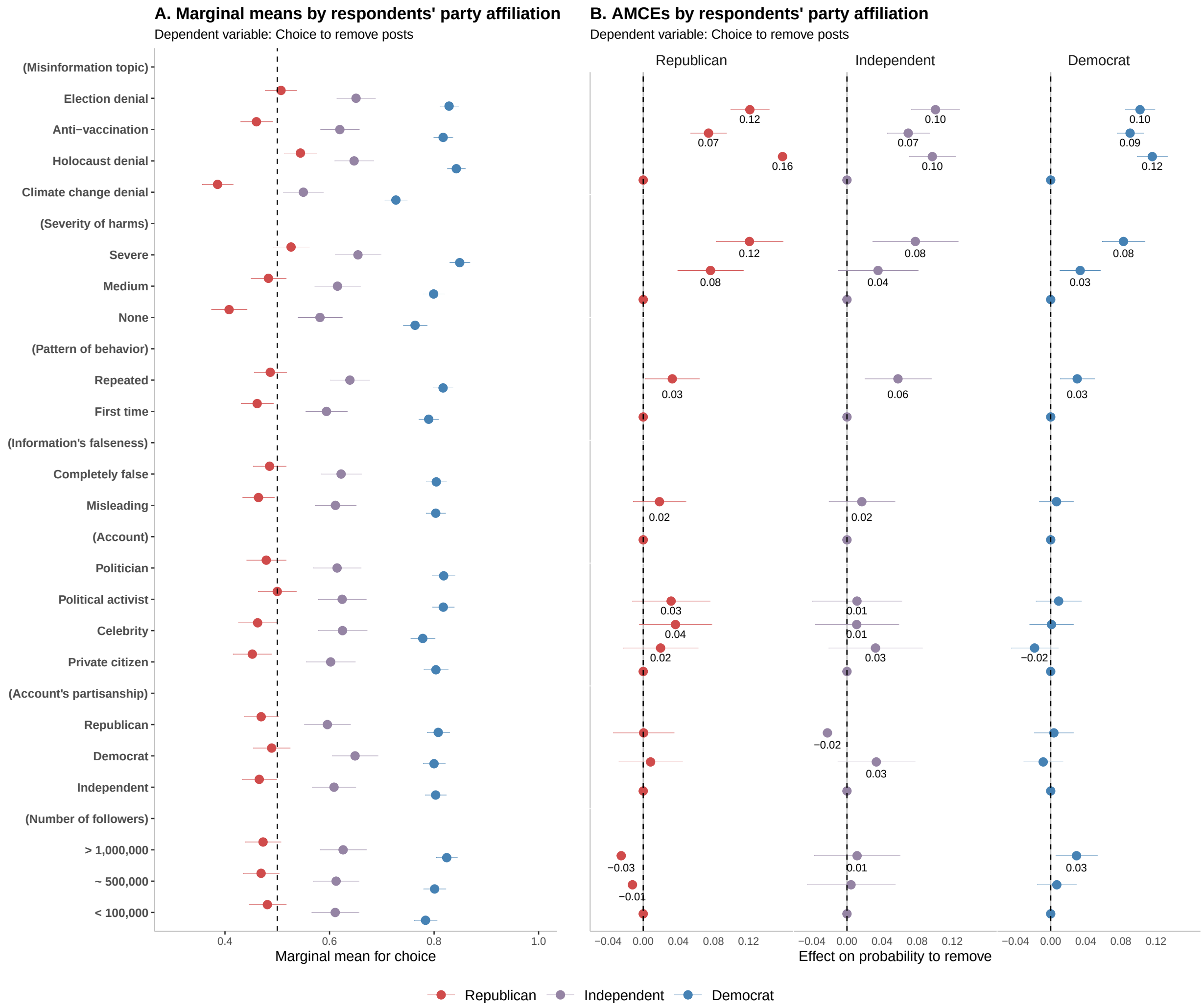


Figure 4: *Respondent subgroup analyses: Differences by political affiliation.* Marginal means point estimates and average marginal component effects (AMCEs) plotted with 95% confidence intervals. Panel A: Marginal means represent the average likelihood of decisions to remove the posts for each attribute level for three respondent subgroups: Republicans, Independents, and Democrats. Dashed line represents the mean value for a binary decision (0.5). Panel B: AMCEs represent effects on probability to remove the posts for each attribute level, faceted by three subgroups: Republicans, Independents, and Democrats. Dashed lines represent the null effect. See Appendix Figure A4 for the subgroup analysis for the rating to penalize accounts.

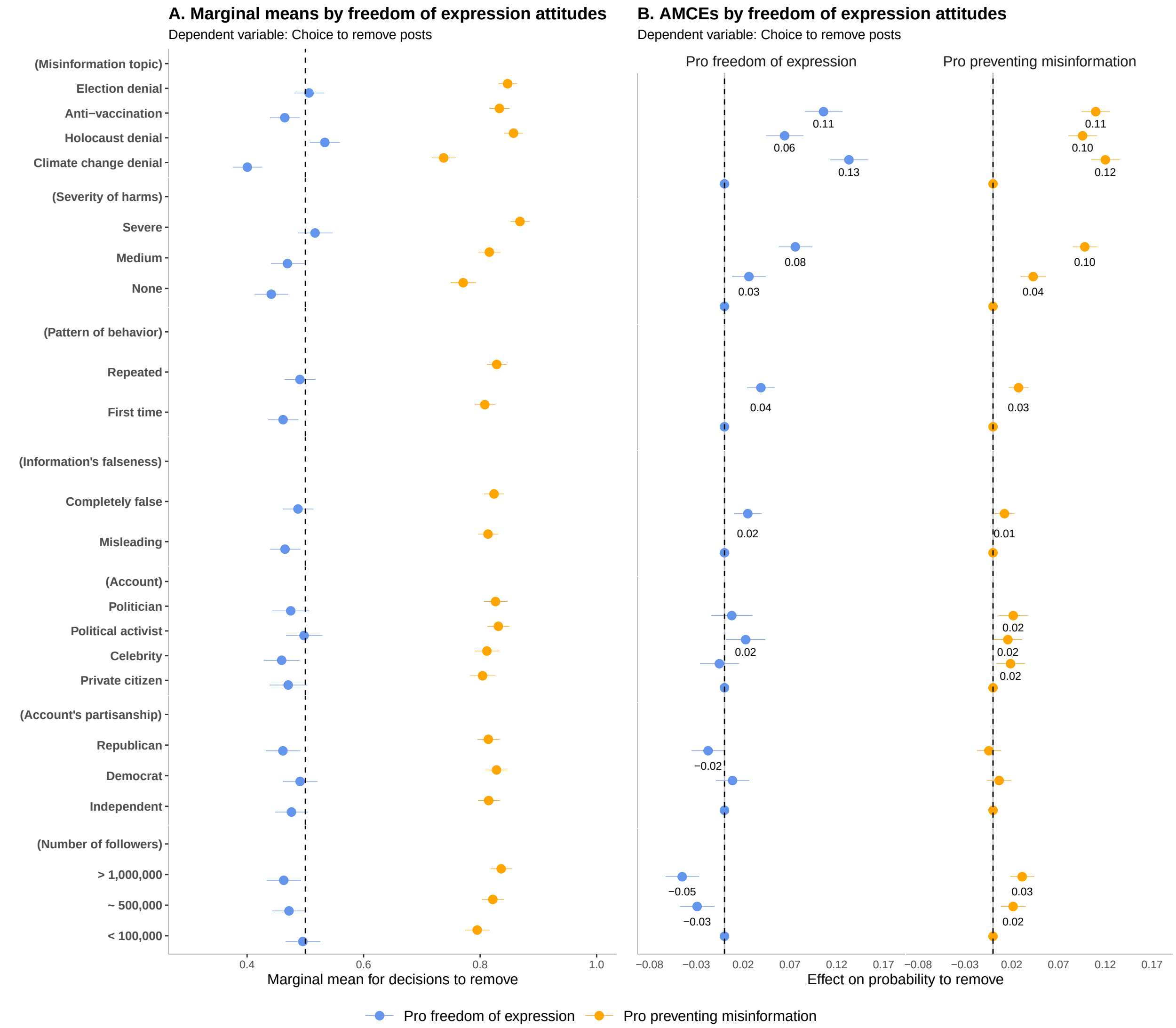


Figure 5: *Respondent subgroup analyses: Differences by attitude toward free speech.* Marginal means point estimates and average marginal component effects (AMCEs) plotted with 95% confidence intervals. Panel A: Marginal means represent the average likelihood of decisions to remove the posts for each attribute level for two respondent subgroups: pro-freedom of expression and pro-mitigating misinformation. Dashed line represents the mean value for a binary decision (0.5). Panel B: AMCEs represent effects on probability to remove the posts for each attribute level, faceted by two subgroups: pro-freedom of expression and pro-mitigating misinformation. Dashed lines represent the null effect. See Appendix Figure A5 for the subgroup analysis for the rating to penalize accounts.

Independents to believe inaccurate claims and disbelieve accurate claims (Appendix Figure B4). For instance, 75% of respondents who described themselves as Democrats indicated that the inaccurate statement "The FDA-approved COVID-19 vaccines can cause infertility" was definitely or possibly false, whereas only 50% of Republicans did. The most contested statement—"The 2020 U.S. Presidential election was stolen from Donald Trump"—was correctly rejected by 84% of Democrats but only 32% of Republicans. Similarly, the accurate statement "There is an overwhelming scientific consensus that human activity (e.g., burning fossil fuels) is the leading cause of climate change" was endorsed by 78% of Democrats but only 36% of Republicans. The only notable exception was the statement related to Holocaust denial, where—irrespective of partisanship—only about 5% of respondents rejected the accurate claim "It is a well established historical fact that 6 million Jews died in the Holocaust" as definitely or probably false (Appendix Figure B4). Since this misinformation topic does not differ along partisan lines, it is of particular interest to our analyses. As Figures 4 and 5 show, Holocaust denial was the only scenario in which a majority of respondents in each partisan group and most pro-free speech respondents took action against the post or account.

The finding that Republicans were more likely to endorse inaccurate claims relevant to our scenarios raises an important question: To what extent do partisan differences in content moderation reflect genuine differences in how respondents weighed the moral dilemmas and to what extent do they merely reflect different views on the facts at hand? Assuming that all respondents, irrespective of partisanship, are less likely to remove posts or suspend accounts if they deem the posted content to be truthful, this alone would predict that Republicans would intervene less. To check the plausibility of this alternative explanation for the partisan differences we found, we conducted a set of robustness analyses where we considered responses in the conjoint part of our study only if the respondent reported an accurate belief in the corresponding belief statement (Appendix C). Here it is important to keep in mind that because these analyses are correlational, they do not license causal claims about the effects of the accuracy of respondents' beliefs on their content moderation decisions. Rather, their purpose is to challenge our findings on partisanship and content moderation. In this subset of responses (Appendix Figure C1), more respondents opted to remove false and misleading posts and penalize the accounts that spread them compared to the full dataset (Figure 2). The majority of Republicans with accurate beliefs were also more likely than Republicans in the full dataset to take action against online misinformation. However, the main patterns in subgroup differences remained (Appendix Figures C3 and C4), including the finding that Republicans were less likely than Independents and Democrats to take action against misinformation. All the main qualitative findings still held when only considering cases evaluated by respondents who endorsed accurate claims relevant to the scenarios. That is, the partisan differences in beliefs only partially accounted for preferences on content moderation and thus do not constitute a viable explanation for the partisan differences in content moderation we observed.

Discussion

Content moderation is controversial and consequential. Regulators are reluctant to restrict harmful but legal content such as misinformation, thereby leaving platforms to decide what content to allow and what to ban. At the heart of policy approaches to online content moderation are trade-offs between fundamental values such as freedom of expression and public health. In our investigation of which aspects of content moderation dilemmas affect people's choices about these trade-offs (i.e., willingness to remove posts and penalize accounts) and what impact individual attitudes have on these decisions, we found that respondents' willingness to remove posts or to suspend an account increased with the severity of the consequences of misinformation. Repeated or habitual offense mattered as well: If the account had previously posted misinformation, respondents were more likely to remove the account. The misinformation topic also mattered——climate change denial was removed the least, whereas Holocaust denial and election denial were removed most often, closely followed by anti-vaccination content. In contrast, features of the account itself—the person behind the account, their partisanship, and number of followers—had little to no effect on respondents' decisions. In other words, the individual characteristics of those who spread misinformation matter little, whereas amount of harm, repeated offense, and type of content matter most.

For the majority of respondents, upholding free speech did not outweigh the detrimental consequences of misinformation. Generally speaking, these results provide support for a consequentialist approach to content moderation of online misinformation. Consequentialism judges the moral permissibility of actions based on their outcomes (Sinnott-Armstrong, 2021). In utilitarianism, a paradigmatic case of moral consequentialism, maximizing happiness (classical utilitarianism; e.g., Bentham, 1781/2000) and minimizing harms for most people (negative utilitarianism; e.g. Smart, 1958) are key ethical principles. Notably, minimizing harm is also one of the most universal ethical principles (e.g., Graham et al., 2013). Results of our study support the idea that this principle also holds for online content moderation. As Twitter's internal survey shows, people support penalties for harmful content online: More than 90% of an international sample supported removing misleading and altered content when it clearly was intended to cause certain types of harm, and more than 75% believed that accounts sharing false and misleading information should be punished (e.g., by having their Tweets deleted or their account suspended; Roth and Achuthan, 2020).

Repeated offense can be classified as *character evidence*—that is, evidence that suggests that a person is likely or not to have acted a certain way based on their reputation, prior conduct, or criminal history. According to our results, repeated sharing of misinformation was a crucial factor in respondents' decisions to remove posts and penalize accounts. Repeated offenses can signal malicious intent, which in turn lends support to the idea that people tend to penalize misinformation shared with malicious intent more than misinformation that might have been shared unwittingly. It is also possible that people are inclined to punish repeated sharing of falsehoods because they consider the potential amplification of harm brought about by repeated sharing (e.g., due to increased exposure to false claims).

Another relevant feature was the number of followers an account had, which is a strong determinant of its reach. Although this feature mattered little on the aggregate level, it was important to different subgroups: Republicans and respondents who were pro-freedom of expression were less likely to penalize accounts with many followers, whereas Democrats and respondents who were pro-mitigating misinformation were more likely to do so. This has interesting implications for the argument of "freedom of reach" (Cohen, 2019), showing, for instance, that accounts with more reach (over 1,000,000 followers in our scenarios) are thought to deserve more protection by respondents who value freedom of expression over mitigating harmful misinformation. Those who prefer to mitigate harmful misinformation, in contrast, appear to be less permissive of accounts with greater reach.

Partisan differences played a major role in people's decisions on content moderation. Respondents did not penalize political out-group accounts more than in-group accounts, but Republicans and Democrats did, in general, make different trade-offs to resolve the dilemma between protecting free speech and removing harmful misinformation. Democrats showed a stronger preference for preventing dangerous falsehoods across all four scenarios, whereas Republicans preferred to protect free speech and imposed fewer restrictions. This partisan divide is consistent with other surveys showing stark partisan divisions in attitudes towards the role of governments and tech firms in restricting online misinformation (Mitchell & Walker, 2021). According to a Pew Research study (Mitchell & Walker, 2021), these differences emerged between 2018 and 2021.

Partisan differences in attitudes toward freedom of expression could be rooted in different approaches to choice autonomy. Republicans' views are likely to be rooted in libertarian philosophy, where individual rights and autonomy are primary values.

Democrats' views, however, are likely to be rooted in a modern liberalism that prioritizes social justice, such that individual rights can be limited for the benefit of society as a whole (see Courtland et al., 2022). These differences in political philosophy might help account for differences between Republicans' and Democrats' attitudes toward removing harmful misinformation online.

Another factor that might account for partisan differences relates to differences in beliefs about the facts at hand. Conservative ideology has been shown to be predictive of endorsement of conspiracy theories among U.S. respondents (van der Linden et al., 2021). Our study also revealed significant partisan divides in respondents' beliefs. Only in the Holocaust denial scenario did beliefs converge across all three partisan subgroups. However, our robustness checks showed that partisan differences remained even when considering only respondents with accurate beliefs about the relevant background knowledge in a scenario (e.g., who correctly dismissed a claim such as "The FDA-approved COVID-19 vaccines can cause infertility"). Partisan differences in beliefs about the facts at hand do not fully explain the partisan differences in content moderation and only partially account for decisions on content moderation.

Given the extent of political polarization in the United States (see, e.g., Klein, 2020), it would have been surprising if Democrats, Republicans, and Independents had uniformly supported the same content moderation measures. And yet, in the majority of cases across the four scenarios, respondents in our study chose to remove the posts and to penalize the offending account. For instance, in the election denial scenario, 49% of respondents chose to temporarily or indefinitely suspend the account, and 31% chose to issue a warning. Assuming that an unheeded warning will eventually be followed by temporary or indefinite suspension, this response pattern implies that even in this highly contentious issue, 80% of respondents prefer taking action over doing nothing. This is particularly important to note in light of the fact that Elon Musk, who is currently in the process of acquiring Twitter, has stated that he would reverse Twitter's permanent ban of Donald Trump, who exemplifies all the attributes that led our participants to take action against a post or account: repeated offenses and severe consequences stemming from his posted misinformation.

One limitation of our study is that both our scenarios and respondents were based in the United States. We chose to focus on the U.S. context for two reasons. First, free speech protectionism is a distinct feature of American culture and politics, and Americans are more supportive of all forms of freedom of expression than are citizens of other countries (Wike & Simmons, 2015). Second, the current debate around content moderation is mostly centered in the United States and many of the rules are being established by U.S.-based companies. However, irrespective of who makes the rules, content moderation affects people across countries and cultures. Ideally, future studies will focus on many different parts of the world. Another limitation is that in our conjoint experiment we stipulated that a user's actions would lead to a specific consequence. In real life, the consequences of a social media post are much harder to establish. Future research should address the impact of risk and uncertainty. A final limitation is that we focused on one type of content moderation dilemmas: when removing harmful but legal content compromises the right to free speech or, conversely, protecting free speech comes at the cost of social harm. But there are many types of content moderation dilemmas. For instance, policing illegal content (e.g., child pornography) through social media raises a dilemma between public safety and individual privacy (see Petrequin, 2022).

When considering the implications of our results for policy, it is important to keep in mind that in liberal democracies, policy makers are reluctant to regulate harmful but not illegal misinformation at the risk of limiting freedom of expression (e.g., European Commission, 2020; Department for Digital, Culture, Media and Sport and Home Office, 2019). The principle of proportionality requires that harsh measures should only be applied when strictly necessary and that a variety of less intrusive mitigating tools should be implemented as a first line of defense. For example, instead of removing outright false content that harms public welfare (e.g., health misinformation), a range of less intrusive measures can be introduced, including warning labels, fact-checking marks, and other prompts that slow the spread of falsehoods (European Commission, 2018, 2021). However, content moderation of harmful content remains a common practice and requires not only cross-platform policy integration but also transparency and consistency in policy development and implementation. Results such as those presented here can contribute to establishing transparent and consistent rules for content moderation that are generally accepted as legitimate. People's preferences are not the only benchmark for making important trade-offs on content moderation, but ignoring them altogether risks undermining the public's trust in content moderation policies.

Materials and Methods

Sample

An online sample of U.S. participants ($N = 2,564$) was recruited by panel provider Ipsos Insights between October 18th and December 3rd, 2021. The sample was quota-matched to the U.S. general population in terms of age, gender, education, ethnicity, and region of residence, with two exceptions where it proved to be infeasible to fill quotas in the online sample: Hispanics and Latinos (ethnicity quota) and people without a high school education (education quota). See Appendix Table A1 for the demographic distribution of the sample.

To determine the required sample size for our study, we conducted two power calculations: with R package *cjpowR* (Schuessler & Freitag, 2020) and simulation-based with R package *DeclareDesign* (Blair et al., 2019; https://declaredesign.org). We estimated AMCE effect sizes for two types of analyses: Within each scenario, we postulated an expected effect size at 0.05 and for all scenarios combined (where misinformation topic is treated as an additional attribute with four levels) at 0.02. Our power analyses are part of our OSF preregistration at https://osf.io/5g8aq.

Study Design

We used a single-profile conjoint survey experiment (Bansak et al., 2021) to explore what influences people's willingness to remove false and misleading content on social media and to penalize accounts that spread it. In the main study task, participants saw 16 cases each (see Figure 1). After excluding missing values in responses, this amounted to a total of 40,845 random cases (see Table A2; the conjoint design described below yielded 1,728 possible unique cases).

Scenarios

Each scenario represented a moral dilemma between freedom of expression and harm from misinformation. The four scenario types represented four misinformation topics: politics (election denial), health (anti-vaccination), history (Holocaust denial), and environment (climate change denial).

Attributes in the Scenarios

Each scenario included seven attributes: (1) person (i.e., who shared information) referred to as the "Account" in the figures; (2) Person's partisanship ("Account's partisanship" in the figures); (3) number of followers; (4) action ("Misinformation topic" in the figures); (5) level of falseness; (6) pattern of behavior; (7) consequences ("Severity of harms" in the figures). Each attribute had multiple levels (Table 1; for the distribution of attribute levels, see Appendix Table A2).

Outcome Measures

Respondents were asked to imagine that they had to decide whether to remove the posts mentioned in the scenarios and whether to suspend the account that posted them. These questions represent two dependent variables, choice to remove posts and rating to penalize account. For choice to remove posts, respondents were asked "What would you do with the posts?" and could answer "remove the posts" or "do nothing." For rating to penalize

Table 1
Conjoint Table

Attribute	**Levels**	**N levels**
Person (Account)	“an elected politician”, “a political activist”, “a celebrity”, “a private citizen”	4
Person (Account) for the “Election denial” scenario	“a presidential candidate”, “a political activist”, “a celebrity”, “a private citizen”	4
Person’s partisanship (Account’s partisanship)	“who is a Democrat”, “who is a Republican”, “who is non-partisan”	3
N of followers	“with less than 100,000 followers on a popular social media platform,” “with about 500,000 followers on a popular social media platform,” “with more than 1 million followers on a popular social media platform,”	3
Action (Misinformation topic) 1 (“Election denial” scenario)	“published a series of posts denying the outcome of the presidential election, encouraging people to join a protest rally and praising violent supporters.”	1
Action (Misinformation topic) 2 (“Anti-vaccination” scenario)	“published a series of posts about serious side effects of the approved COVID-19 vaccines (e.g., that vaccines cause infertility).”	1
Action (Misinformation topic) 3 (“Holocaust denial” scenario)	“published a series of posts questioning the scale of the Holocaust (e.g., that significantly fewer than 6 million Jews were killed).”	1
Action (Misinformation topic) 4 (“Climate change denial” scenario)	“published a series of posts denying scientific consensus that human activity (e.g., burning fossil fuels) is the leading cause of climate change.”	1
Level of falseness	“The specific information they shared is completely false and negates the established facts.”; “The specific information they shared is misleading and distorts the established facts.”	2
Pattern of behavior	“This was the first time they shared false or misleading information.”, “This was not the first time they shared false or misleading information.”	2
Consequences (Severity of harms) 1 (“Election denial” scenario)	No consequences: “Suppose you know that these messages caused no consequences.”; Medium: “Suppose you know that, due to this, a nonviolent demonstration occurred.”; Severe: “Suppose you know that, due to this, a violent demonstration occurred, 5 people died, and 150 protesters were detained.”	3
Consequences (Severity of harms) 2 (“Anti-vaccination” scenario)	“Suppose you know that these messages caused no consequences.”, “Suppose you know that, due to this, 10,000 citizens who were planning to get a vaccine refused to vaccinate.”, “Suppose you know that, due to this, 1 million people who were planning to get a vaccine refused to vaccinate, resulting in approximately 10,000 additional deaths.”	3
Consequences (Severity of harms) 3 (“Holocaust denial” scenario)	“Suppose you know that these messages caused no consequences.”, “Suppose you know that, due to this, several antisemitic attacks occurred, with no severe injuries.”, “Suppose you know that, due to this, several antisemitic attacks occurred, injuring 2 people and killing 1 person.”	3
Consequences (Severity of harms) 4 (“Climate change denial” scenario)	“Suppose you know that these messages caused no consequences.”, “Suppose you know that these posts convinced 1,000 people that climate change is a hoax.”, “Suppose you know that these posts convinced 100,000 voters that climate change is a hoax, thereby swinging the outcome of the next election and preventing the passage of a bill that would have cut carbon emissions by 20%.”	3

account, respondents were asked “What would you do with this user’s account?” and could answer “suspend the account indefinitely,” “suspend the account temporarily,” “issue a warning,” or “do nothing” (see Figure 1). Each participant saw 16 scenario variations (four variations of four scenario types) and gave two responses for each (32 responses in total).

Attention Check

A simple attention check was presented at the start of the study: Participants were asked, “How many scenarios are you expected to see?” The question was displayed on the same page as the description of the main task, which included the correct answer (16) in bold characters. Participants who did not pass the attention check were redirected to the study termination page. This information was included in the consent form.

Demographics and Political Attitudes

After giving informed consent, and prior to the main study task, respondents filled out demographic information and information on their political attitudes (Table A1).

Perceived Accuracy, Harm, and Severity of Outcomes

After the main study task, respondents were asked to rate the accuracy of four statements, each relevant to a different scenario (“The 2020 U.S. Presidential election was stolen from Donald Trump”, ‘The FDA-approved COVID-19 vaccines can cause infertility”,“Death of 6 million Jews in the Holocaust is a well established historical fact”, “There is an overwhelming scientific consensus that human activity (e.g., burning fossil fuels) is the leading cause of climate change”) on a 5-point Likert scale (definitely false, probably false, don’t know, probably true, definitely true; Figure B4). They were also asked to rate the perceived harm of the content featured in each scenario on a 5-point Likert scale (not at all harmful, a little, somewhat, very, extremely harmful; Figure B5) and the perceived severity of the outcomes featured in each scenario on a 5-point Likert scale (not severe at all, slightly, somewhat, very, extremely severe; Appendix Figure B6).

Attitudes Toward Freedom of Expression

We included two sets of measures of people's attitudes toward freedom of expression and its limitations. First, four questions addressed participants' general attitudes toward freedom of expression and its limits in cases of prejudice, falsehoods, and potential for harm (four items adapted from Riedl et al., 2021; for items and distribution of responses, see Appendix Figure B3). Second, two questions addressed people's preferences in the dilemma between freedom of expression and preventing harmful misinformation: one on the relative importance of freedom of expression versus preventing disinformation from spreading and another on platform choice to choose between a hypothetical social media platform that always prioritizes free speech and another that moderates content strictly. Participants answered these two questions both before and after the main study task so that we could compare proportions of respondents who were willing to impose limits on free expression to mitigate harmful misinformation before and after they faced the moral dilemmas in our scenarios (for items and distribution of responses, see Appendix Figure B2).

Estimates of N of Misinformation Factors

We administered one item after the second set of questions on attitudes toward freedom of expression. This item asked participants to estimate how many accounts produce the majority of misinformation on social media ("To the best of your knowledge, how many individuals are responsible for 65% of the anti-vaccination disinformation on Facebook and Twitter? Please indicate or estimate a number."). We based the correct answer on the Center for Countering Digital Hate's (2021) recent estimate that 12 accounts are responsible for 65% of the anti-vaccination misinformation on Facebook and Twitter. For results, see Appendix Figure B7.

The full study instrument is available on OSF at https://osf.io/2s4vn/.

Data Analysis

We employed a combination of descriptive and inferential statistics. For descriptive analyses, we reported the demographic distribution of our sample, frequencies of conjoint features, and proportions of choices for several measures in the study.

The main analysis was the conjoint analysis used to estimate casual effects of multiple factors (attributes) on the binary decision to remove posts in all four scenarios and the rating measure on whether to suspend an account (permanently or temporarily), issue a warning, or do nothing.

We conducted the main analysis using *cregg* (Leeper, 2020), an R package for analyzing and visualizing the results of conjoint experiments. Although we had initially intended to use *cjoint* (Hainmueller, Hopkins, et al., 2014), we deviated from our preregistration due to *cregg*'s superior functionality for our purposes. We reported on estimates for the following estimands (see Bansak et al., 2021; Hainmueller, Hopkins, et al., 2014; Leeper et al., 2020): marginal means and AMCEs. Marginal means facilitate interpretations of conjoint attributes' impact on respondents' decisions not predicated on a specific reference category, whereas AMCEs show effect sizes relative to the chosen reference levels (see Leeper et al., 2020).

Preregistration

The study was preregistered at OSF (https://osf.io/5g8aq). The preregistration also includes the full study instrument and the power analysis.

Analyses of all measures included in the study and of some preregistered research questions that did not appear in the main text are provided in Appendix B. These additional results do not alter any of the results and conclusions we present in the main text.

Ethics

Informed consent was obtained from all participants and the study was conducted in accordance with relevant guidelines and regulations. The Institutional Review Board of the Max Planck Institute for Human Development approved the study (approval C2021-16).

Data Availability

Anonymized data and code are available at OSF (https://osf.io/2s4vn/).

Supporting Information

The supporting information consists of four appendices. Appendix A includes supplementary figures and tables for methods and conjoint analyses. Appendix B includes visualizations of descriptive and summary statistics for study measures that complement our main analyses. Appendix C includes the subset analysis for our outcome variables based on the accuracy of the relevant background knowledge. Appendix D includes an overview table of misinformation policies for major social media platforms.

Authors' Contributions

Conceptualization: A.K., S.L., R.H., S.M.H., P.L.-S., M.L., and J.R. Data curation: A.K. Formal analysis: A.K., S.M.H., and J.R. Funding acquisition: S.L., R.H., S.M.H., and P.L.-S. Investigation: A.K. Methodology: A.K., S.L., R.H., S.M.H., P.L.-S., and J.R. Project administration: A.K. Software: A.K. and J.R. Supervision: S.M.H., S.L., R.H., and J.R. Visualization: A.K. and S.M.H. Writing - original draft: A.K. Writing - review editing: A.K., S.L., R.H., S.M.H., P.L.-S., M.L., and J.R.

Correspondence concerning this article should be addressed to Anastasia Kozyreva, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin. Email: kozyreva@mpib-berlin.mpg.de

Funding and Acknowledgements

The study was funded by the grant of the Volkswagen Foundation to R.H., S.L., and S.M.H. (Initiative “Artificial Intelligence and the Society of the Future”). S.L. was supported by a Research Award from the Humboldt Foundation in Germany while this research was conducted. S.L. also acknowledges financial support from the European Research Council (ERC Advanced Grant 101020961 PRODEMINFO).

The authors thank the Ipsos Insights team for their help with data collection. We are also grateful to Spela Vrtovec for research assistance, to Deb Ain for editing the manuscript, and to our colleagues at the Center for Adaptive Rationality for their feedback and productive discussions.

Competing Interests

The authors declare no competing interests.

References

Abrams v. United States, 250 U.S. 616 (1919).

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A. C., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., J, M., Opoku-Agyemang, K., Scheich Borg, J., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., & Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences*, *26*(5). https://doi.org/10.1016/j.tics.2022.02.009

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2021). Conjoint survey experiments. In J. N. Druckman & D. P. Green (Eds.), *Advances in experimental political science*. Cambridge University Press. https://doi.org/10.1017/9781108777919.004

Bentham, J. (2000). *An introduction to the principles of morals and legislation*. Batoche Books. (Original work published 1781)

Bickert, M. (2020, October 12). *Removing Holocaust denial content*. Meta. https://about.fb.com/news/2020/10/removing-holocaust-denial-content/

Blair, G., Cooper, J., Coppock, A., & Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, *113*(3), 838–859. https://doi.org/10.1017/S0003055419000194

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Clegg, N. (2021, June 4). *In response to Oversight Board, Trump suspended for two years; will only be reinstated if conditions permit*. Meta.

https://about.fb.com/news/2021/06/facebook-response-to-oversight-board-recommendations-trump/

Cohen, S. B. (2019, November 21). *Sacha Baron Cohen's keynote address at ADL's 2019 Never Is Now summit on anti-Semitism and hate*. ADL. https://www.adl.org/news/article/sacha-baron-cohens-keynote-address-at-adls-2019-never-is-now-summit-on-anti-semitism

Courtland, S. D., Gaus, G., & Schmidtz, D. (2022). Liberalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2022). Stanford University. https://plato.stanford.edu/archives/spr2022/entries/liberalism/

Department for Digital, Culture, Media and Sport, & Home Office. (2019). *Online harms white paper*. Crown. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973939/Online_Harms_White_Paper_V2.pdf

Douek, E. (2021). Governing online speech. *Columbia Law Review*, *121*(3), 759–834. https://doi.org/10.2139/ssrn.3679607

European Commission. (2018). *Code of practice on disinformation*. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

European Commission. (2020). Proposal for a regulation of the European Parliament and of the Council on a single market for digital services (Digital Services Act) and amending Directive 2000/31/EC (COM/2020/825 final). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en

European Commission. (2021). European Commission guidance on strengthening the Code of Practice on Disinformation (COM(2021)262 final). https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Funke, D., & Flamini, D. (n.d.). *A guide to anti-misinformation actions around the world.* Poynter. Retrieved January 27, 2022, from https://www.poynter.org/ifcn/anti-misinformation-actions/.

Google. (n.d.-a). *COVID-19 medical misinformation policy.* Retrieved December 26, 2021, from https://support.google.com/youtube/answer/9891785?hl=en.

Google. (n.d.-b). *Misrepresentation.* Retrieved December 27, 2021, from https://support.google.com/adspolicy/answer/6020955?hl=en.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (pp. 55–130). Academic Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Hainmueller, J., Hopkins, D., & Yamamoto, T. (2014). *cjoint: Causal inference in conjoint analysis: Understanding multi-dimensional choices via stated preference experiments* [R package version 2.1.0]. https://CRAN.R-project.org/package=cjoint

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, *22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

Instagram. (n.d.). *COVID-19 and vaccine policy updates and protections.* Retrieved December 27, 2021, from https://help.instagram.com/697825587576762.

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, *345*(6199), Article 1251722. https://doi.org/10.1126/science.1251722

Klein, E. (2020). *Why we're polarized.* Simon & Schuster.

Leeper, T. J. (2020). *Cregg: Simple conjoint analyses and visualization* [R package version 0.4.0].

Leeper, T. J., Hobolt, S. B., & Tilley, J. (2020). Measuring subgroup preferences in conjoint experiments. *Political Analysis*, *28*(2), 207–221. https://doi.org/10.1017/pan.2019.30

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

Mill, J. S. (2011). *On liberty.* Cambridge University Press. (Original work published 1859)

Mitchell, A., & Walker, M. (2021, August 18). *More Americans now say government should take steps to restrict false information online than in 2018.* Pew Research Center. https://www.pewresearch.org/fact-tank/2021/08/18/more-americans-now-say-government-should-take-steps-to-restrict-false-information-online-than-in-2018/

Monbiot, G. (2021). Covid lies cost lives – we have a duty to clamp down on them. *The Guardian.* https://www.theguardian.com/commentisfree/2021/jan/27/covid-lies-cost-lives-right-clamp-down-misinformation

Petrequin, S. (2022). EU commission proposes plan to fight child pornography. *The Washington Post.* https://www.washingtonpost.com/world/eu-commission-proposes-plan-to-fight-child-pornography/2022/05/11/be8991f6-d11a-11ec-886b-df76183d233f_story.html

Pinterest. (n.d.). *Community guidelines.* Retrieved April 26, 2022, from https://policy.pinterest.com/en/community-guidelines.

Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet*, *13*(3), 433–451. https://doi.org/https://doi.org/10.1002/poi3.257

Rosen, G. (2020, April 16). An update on our work to keep people informed and limit misinformation about COVID-19. *Meta*. Retrieved January 4, 2022, from https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims.

Roth, Y., & Achuthan, A. (2020, February 4). *Building rules in public: Our approach to synthetic manipulated media*. Twitter. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media

Schauer, F. (1978). Fear, risk and the First Amendment: Unraveling the chilling effect. *Boston University Law Review*, *58*, 685–732.

Schuessler, J., & Freitag, M. (2020). *Power analysis for conjoint experiments* [Unpublished manuscript]. https://doi.org/10.31235/osf.io/9yuhp

Sinnott-Armstrong, W. (2021). Consequentialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021). Stanford University. https://plato.stanford.edu/archives/fall2021/entries/consequentialism/

Sinott-Armstrong, W. (1988). *Moral dilemmas*. Blackwell.

Smart, R. (1958). Negative utilitarianism. *Mind*, *LXVII*(268), 542–543. https://doi.org/10.1093/mind/lxvii.268.542

Sunstein, C. R. (2021). *Liars: Falsehoods and free speech in an age of deception*. Oxford University Press. https://doi.org/10.1093/oso/9780197545119.001.0001

The Economist. (2020, October 20). Social media's struggle with self-censorship. https://www.economist.com/briefing/2020/10/22/social-medias-struggle-with-self-censorship

The Economist. (2021, February 13). *Censorious governments are abusing "fake news" laws.* https://www.economist.com/international/2021/02/11/censorious-governments-are-abusing-fake-news-laws

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395–1415. https://doi.org/10.2307/796133

Twitter. (2021, January 8). *Permanent suspension of @realDonaldTrump.* https://blog.twitter.com/en_us/topics/company/2020/suspension

Twitter. (n.d.-a). *COVID-19 misleading information policy.* Retrieved December 26, 2021, from https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

Twitter. (n.d.-b). *Hateful conduct policy.* Retrieved December 27, 2021, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

United Nations. (1948). *Universal declaration of human rights.* https://www.un.org/en/about-us/universal-declaration-of-human-rights

van der Linden, S., Panagopoulos, C., Azevedo, F., & Jost, J. T. (2021). The paranoid style in American politics revisited: An ideological asymmetry in conspiratorial thinking. *Political Psychology*, *42*(1), 23–51. https://doi.org/10.1111/pops.12681

van Mill, D. (2021). Freedom of speech. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021). Stanford University. https://plato.stanford.edu/archives/spr2021/entries/freedom-speech/

Wike, R., & Simmons, K. (2015, November 15). *Global support for principle of free expression, but opposition to some forms of speech.* Pew Research Center. https://www.pewresearch.org/global/2015/11/18/global-support-for-principle-of-free-expression-but-opposition-to-some-forms-of-speech/

Wiseman, J. (2020, October 3). *Rush to pass 'fake news' laws during Covid-19 intensifying global media freedom challenges.* International Press Institute. https://ipi.media/rush-to-pass-fake-news-laws-during-covid-19-intensifying-global-media-freedom-challenges/

Yadav, K., Erdoğdu, U., Siwakoti, S., Shapiro, J. N., & Wanless, A. (2021). Countries have more than 100 laws on the books to combat misinformation. How well do they work? *Bulletin of the Atomic Scientists*, *77*(3), 124–128. https://doi.org/10.1080/00963402.2021.1912111

Yang, M. (2022, January 14). *'Menace to public health': 270 experts criticise Spotify over Joe Rogan's podcast*. The Guardian. https://www.theguardian.com/technology/2022/jan/14/spotify-joe-rogan-podcast-open-letter

YouTube Insider [@YoutubeInsider]. (2021, January 13). *After review, and in light of concerns about the ongoing potential for violence, we removed new content uploaded to Donald* [Tweet]. Twitter. https://twitter.com/YouTubeInsider/status/1349205688694812672?s=20&t=pMA3f60oCs6NI5ALZuI-Zw

Zuckerberg, M. (2016, November 19). *A lot of you have asked what we're doing about misinformation, so I wanted to give an update. The bottom* [Status update]. Facebook. https://www.facebook.com/zuck/posts/10103269806149061

Zuckerberg, M. (2019, October 19). *Standing for voice and free expression.* https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/

Appendix A

Methods and Conjoint Analyses

Table A1

Sample Information

Sample size	
N	2564.0
Duration (minutes)	
Duration (mean)	21.0
Duration (median)	13.0
Gender (%)	
Male	48.5
Female	51.2
Other	0.3
Age group (%)	
Age (18–24)	9.4
Age (25–34)	16.9
Age (35–44)	17.1
Age (45–54)	16.7
Age (55–64)	17.6
Age (65+)	22.4
Education (%)	
Less than high school	2.7
High school	27.0
Some college	19.9
Associate or undergraduate degree	34.4
Master's degree	13.1
Doctoral degree	3.0
Region of residency (%)	
South	36.6
West	23.2
Northeast	19.7
Midwest	20.4
Ethnicity (%)	
White	70.8
Black or African-American	12.8
Hispanic or Latino	6.4
Asian or Asian-American	7.2
Other	2.8
Political party (%)	
Democrat	47.3
Independent or not sure	20.1
Republican	32.5
Political ideology (%)	
Liberal	29.6
Moderate or not sure	42.6
Conservative	27.8

Table A2

Frequency of Conjoint Features

Attribute and Levels	N	%
Account		
Private citizen	9,558	23.4
Celebrity	10,789	26.4
Political activist	10,804	26.5
Politician	9,694	23.7
Account's partisanship		
Independent	14,453	35.4
Democrat	12,618	30.9
Republican	13,774	33.7
N of followers		
< 100,000	12,820	31.4
~ 500,000	12,835	31.4
> 1,000,000	15,190	37.2
Action/Misinformation topic		
Climate change denial	10,256	25.1
Holocaust denial	10,077	24.7
Anti-vaccination	10,256	25.1
Election denial	10,256	25.1
Level of falseness		
Misleading	20,957	51.3
Completely false	19,888	48.7
Pattern of behavior		
First time	20,249	49.6
Repeated	20,596	50.4
Consequences/Severity of harms		
None	13,685	33.5
Medium	13,380	32.8
Severe	13,780	33.7
Total N per attribute		
	40,845	

Table A3

AMCEs for choice to remove post

	Attribute	Level	Estimate	SE	Lower CI	Upper CI
1	Misinformation topic	Climate change denial	0.00			
2	Misinformation topic	Holocaust denial	0.13	0.01	0.11	0.14
3	Misinformation topic	Anti-vaccination	0.08	0.01	0.07	0.09
4	Misinformation topic	Election denial	0.11	0.01	0.10	0.12
5	Severity of harms	None	0.00			
6	Severity of harms	Medium	0.04	0.01	0.02	0.06
7	Severity of harms	Severe	0.09	0.01	0.07	0.11
8	Pattern of behavior	First time	0.00			
9	Pattern of behavior	Repeated	0.04	0.01	0.02	0.06
10	Information's falseness	Misleading	0.00			
11	Information's falseness	Completely false	0.02	0.01	0.00	0.03
12	Account	Private citizen	0.00			
13	Account	Celebrity	0.01	0.01	-0.02	0.03
14	Account	Political activist	0.02	0.01	-0.01	0.04
15	Account	Politician	0.02	0.01	-0.01	0.04
16	Account's partisanship	Independent	0.00			
17	Account's partisanship	Democrat	0.01	0.01	-0.01	0.03
18	Account's partisanship	Republican	-0.00	0.01	-0.02	0.02
19	Number of followers	< 100,000	0.00			
20	Number of followers	~ 500,000	-0.01	0.01	-0.03	0.01
21	Number of followers	> 1,000,000	-0.00	0.01	-0.03	0.02

SE = standard error; CI = confidence interval.

Table A4

AMCEs for Rating to Penalize Account

	Attribute	Level	Estimate	SE	Lower CI	Upper CI
1	Misinformation topic	Climate change denial	0.00			
2	Misinformation topic	Holocaust denial	0.36	0.01	0.34	0.39
3	Misinformation topic	Anti-vaccination	0.21	0.01	0.19	0.23
4	Misinformation topic	Election denial	0.36	0.01	0.33	0.38
5	Severity of harms	None	0.00			
6	Severity of harms	Medium	0.15	0.02	0.11	0.19
7	Severity of harms	Severe	0.41	0.02	0.36	0.45
8	Pattern of behavior	First time	0.00			
9	Pattern of behavior	Repeated	0.30	0.02	0.26	0.34
10	Information's falseness	Misleading	0.00			
11	Information's falseness	Completely false	0.04	0.02	0.01	0.08
12	Account	Private citizen	0.00			
13	Account	Celebrity	0.02	0.02	-0.03	0.07
14	Account	Political activist	0.05	0.02	0.01	0.10
15	Account	Politician	0.05	0.03	0.00	0.10
16	Account's partisanship	Independent	0.00			
17	Account's partisanship	Democrat	0.03	0.02	-0.01	0.07
18	Account's partisanship	Republican	0.02	0.02	-0.02	0.06
19	Number of followers	< 100,000	0.00			
20	Number of followers	~ 500,000	0.03	0.02	-0.01	0.07
21	Number of followers	> 1,000,000	0.03	0.02	-0.02	0.07

SE = standard error; CI = confidence interval.

Table A5

Marginal Means for Choice to Remove Post

	Attribute	Level	Estimate	SE	Lower CI	Upper CI
1	Misinformation topic	Climate change denial	0.58	0.01	0.56	0.60
2	Misinformation topic	Holocaust denial	0.71	0.01	0.69	0.72
3	Misinformation topic	Anti-vaccination	0.66	0.01	0.64	0.68
4	Misinformation topic	Election denial	0.69	0.01	0.67	0.70
5	Severity of harms	None	0.61	0.01	0.60	0.63
6	Severity of harms	Medium	0.66	0.01	0.64	0.67
7	Severity of harms	Severe	0.70	0.01	0.68	0.72
8	Pattern of behavior	First time	0.64	0.01	0.62	0.66
9	Pattern of behavior	Repeated	0.67	0.01	0.66	0.69
10	Information's falseness	Misleading	0.65	0.01	0.63	0.67
11	Information's falseness	Completely false	0.67	0.01	0.65	0.68
12	Account	Private citizen	0.65	0.01	0.63	0.67
13	Account	Celebrity	0.65	0.01	0.63	0.67
14	Account	Political activist	0.67	0.01	0.65	0.69
15	Account	Politician	0.67	0.01	0.65	0.69
16	Account's partisanship	Independent	0.65	0.01	0.63	0.67
17	Account's partisanship	Democrat	0.67	0.01	0.65	0.69
18	Account's partisanship	Republican	0.65	0.01	0.64	0.67
19	Number of followers	< 100,000	0.66	0.01	0.64	0.68
20	Number of followers	~ 500,000	0.65	0.01	0.64	0.67
21	Number of followers	> 1,000,000	0.66	0.01	0.64	0.68

SE = standard error; CI = confidence interval.

Table A6

Marginal Means for Rating to Penalize Account

	Attribute	Level	Estimate	SE	Lower CI	Upper CI
1	Misinformation topic	Climate change denial	2.18	0.02	2.15	2.22
2	Misinformation topic	Holocaust denial	2.55	0.02	2.51	2.58
3	Misinformation topic	Anti-vaccination	2.39	0.02	2.36	2.43
4	Misinformation topic	Election denial	2.54	0.02	2.51	2.58
5	Severity of harms	None	2.24	0.02	2.20	2.28
6	Severity of harms	Medium	2.38	0.02	2.34	2.42
7	Severity of harms	Severe	2.62	0.02	2.58	2.67
8	Pattern of behavior	First time	2.28	0.02	2.24	2.31
9	Pattern of behavior	Repeated	2.55	0.02	2.51	2.59
10	Information's falseness	Misleading	2.41	0.02	2.38	2.45
11	Information's falseness	Completely false	2.42	0.02	2.38	2.45
12	Account	Private citizen	2.36	0.02	2.32	2.41
13	Account	Celebrity	2.36	0.02	2.32	2.40
14	Account	Political activist	2.48	0.02	2.44	2.53
15	Account	Politician	2.45	0.02	2.41	2.49
16	Account's partisanship	Independent	2.38	0.02	2.34	2.42
17	Account's partisanship	Democrat	2.44	0.02	2.40	2.48
18	Account's partisanship	Republican	2.44	0.02	2.39	2.48
19	Number of followers	< 100,000	2.37	0.02	2.33	2.41
20	Number of followers	~ 500,000	2.42	0.02	2.38	2.46
21	Number of followers	> 1,000,000	2.45	0.02	2.41	2.49

SE = standard error; CI = confidence interval.

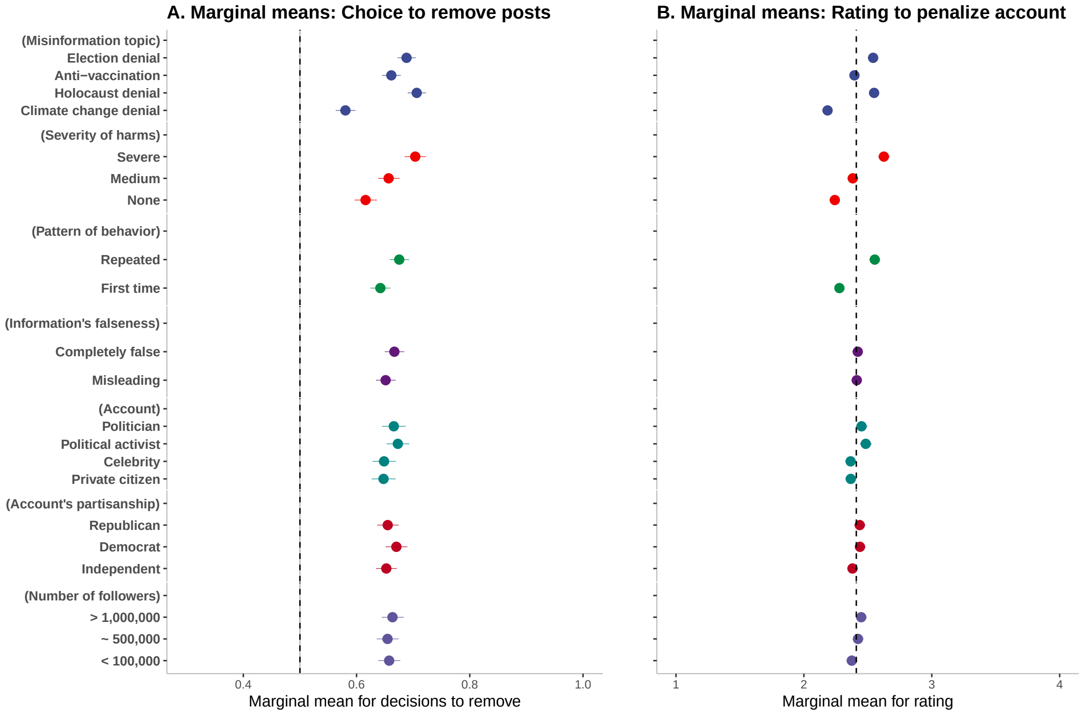


Figure A1: *Marginal means for decisions across all scenarios.* Marginal means point estimates plotted with 95% CIs. Panel A: Marginal means represent the average likelihood of decisions to remove the posts for each attribute level and are shown relative to the mean value of 0.5. Values above the dashed line indicate increased likelihood to remove relative to equal likelihood. Panel B: Marginal means represent the average rating to penalize the account for each attribute level and are shown relative to the grand mean (i.e., overall mean for all attribute levels = 2.4 on a 4-point scale, where 1 is "do nothing" and 4 "indefinitely suspend"). Values above the dashed line indicate levels that increase rating to penalize account relative to the grand mean. For all marginal means estimates, see Appendix Tables A5, A6.

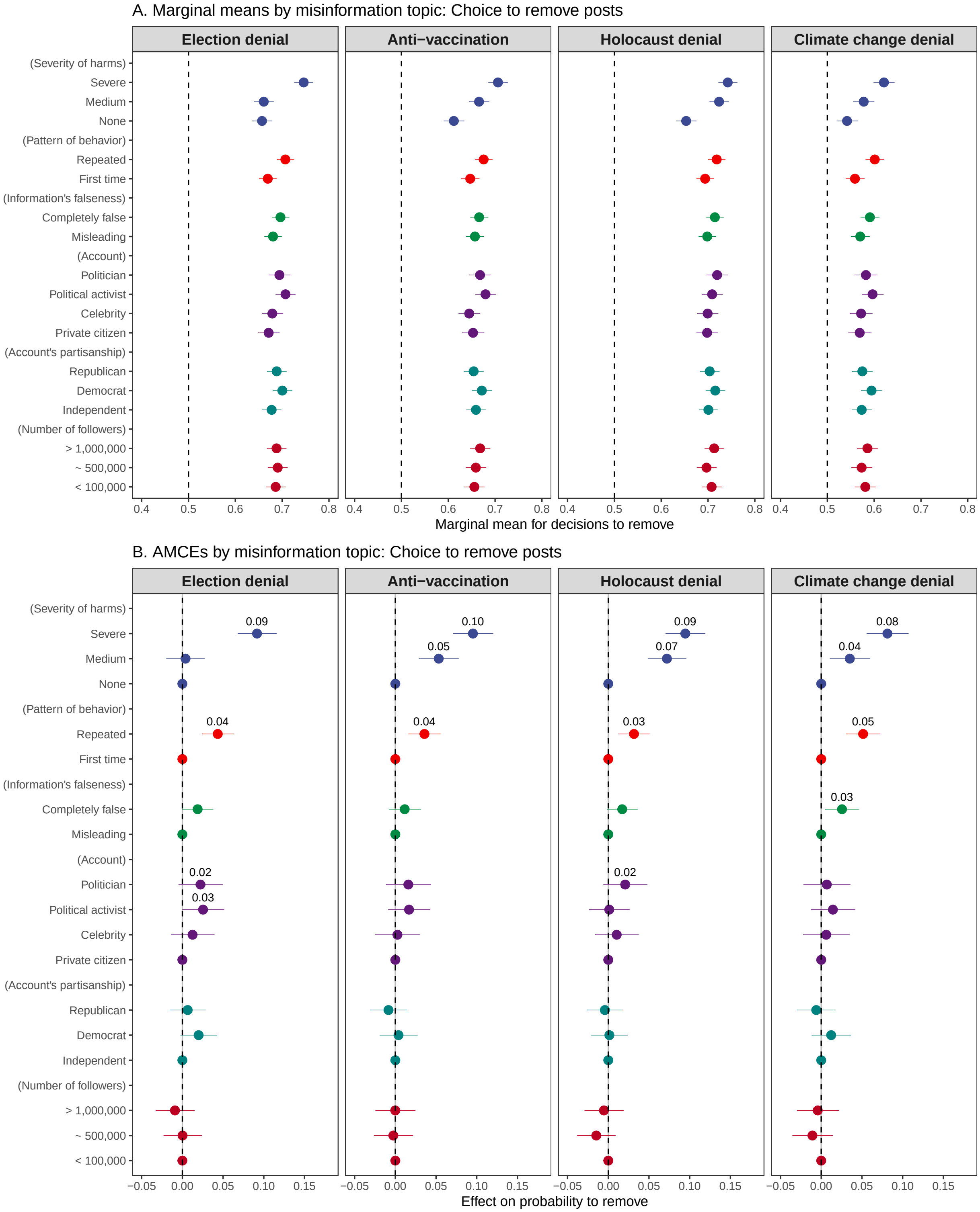


Figure A2: *Marginal means and average marginal component effects (AMCEs) for choices to remove the post for each misinformation topic.* Marginal means point estimates and AMCEs plotted with 95% CIs. Panel A: Marginal means represent the average likelihood of decisions to remove the posts for each attribute level faceted by four misinformation topics. Dashed lines represent the mean value for a binary decision (0.5). Panel B: AMCEs represent effects on probability to remove the posts for each attribute level faceted by four misinformation topics. Dashed lines represent the null effect.

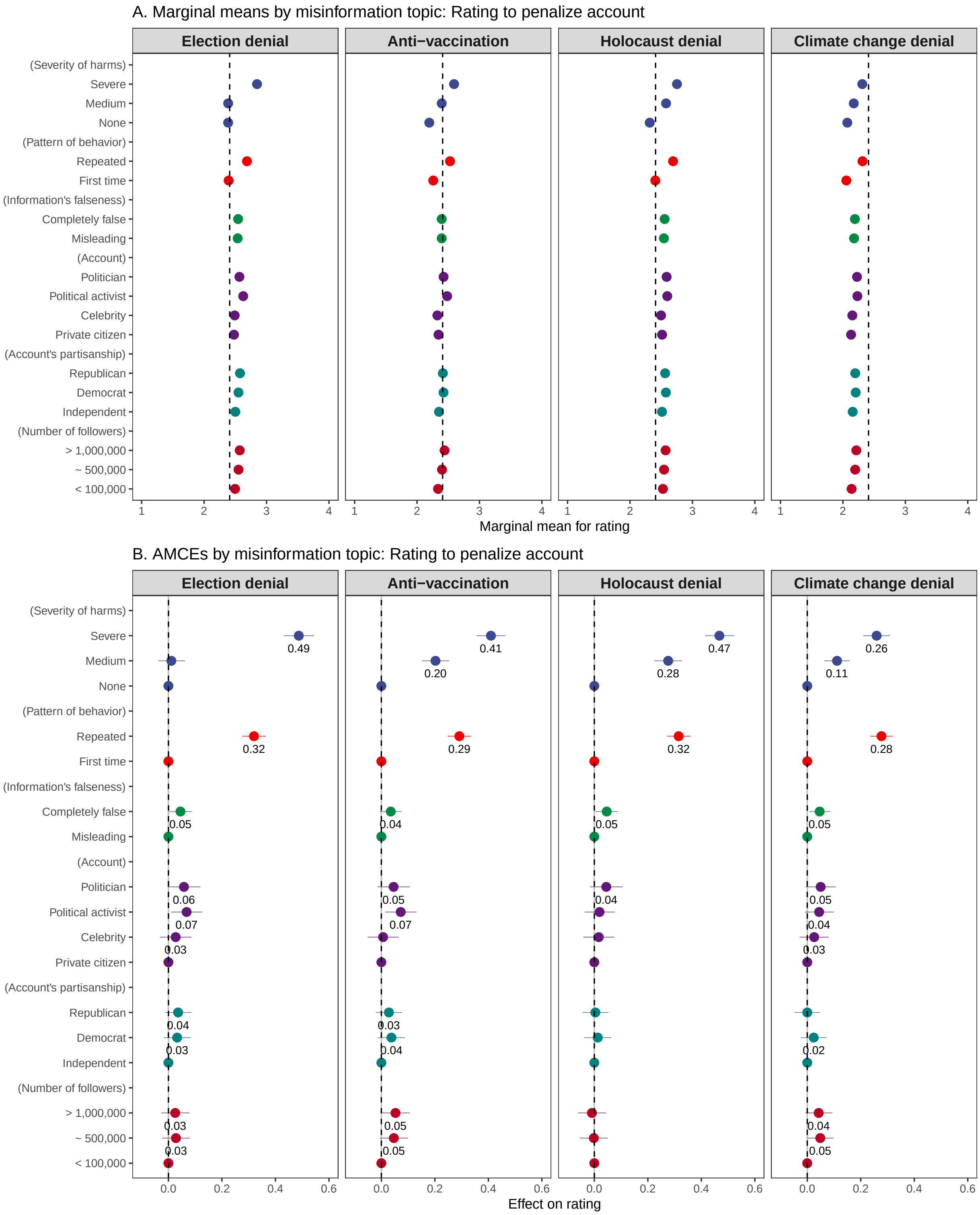


Figure A3: *Marginal means and average marginal component effects (AMCEs) for ratings to penalize accounts for each misinformation topic.* Marginal means point estimates and AMCEs plotted with 95% CIs. Panel A: Marginal means represent the average rating for decisions to penalize the account for each attribute level, faceted by four misinformation topics. Dashed lines represent grand mean for rating (2.41). Panel B: AMCEs represent effects on rating to penalize the account for each attribute level, faceted by four misinformation topics. Dashed lines represent the null effect.

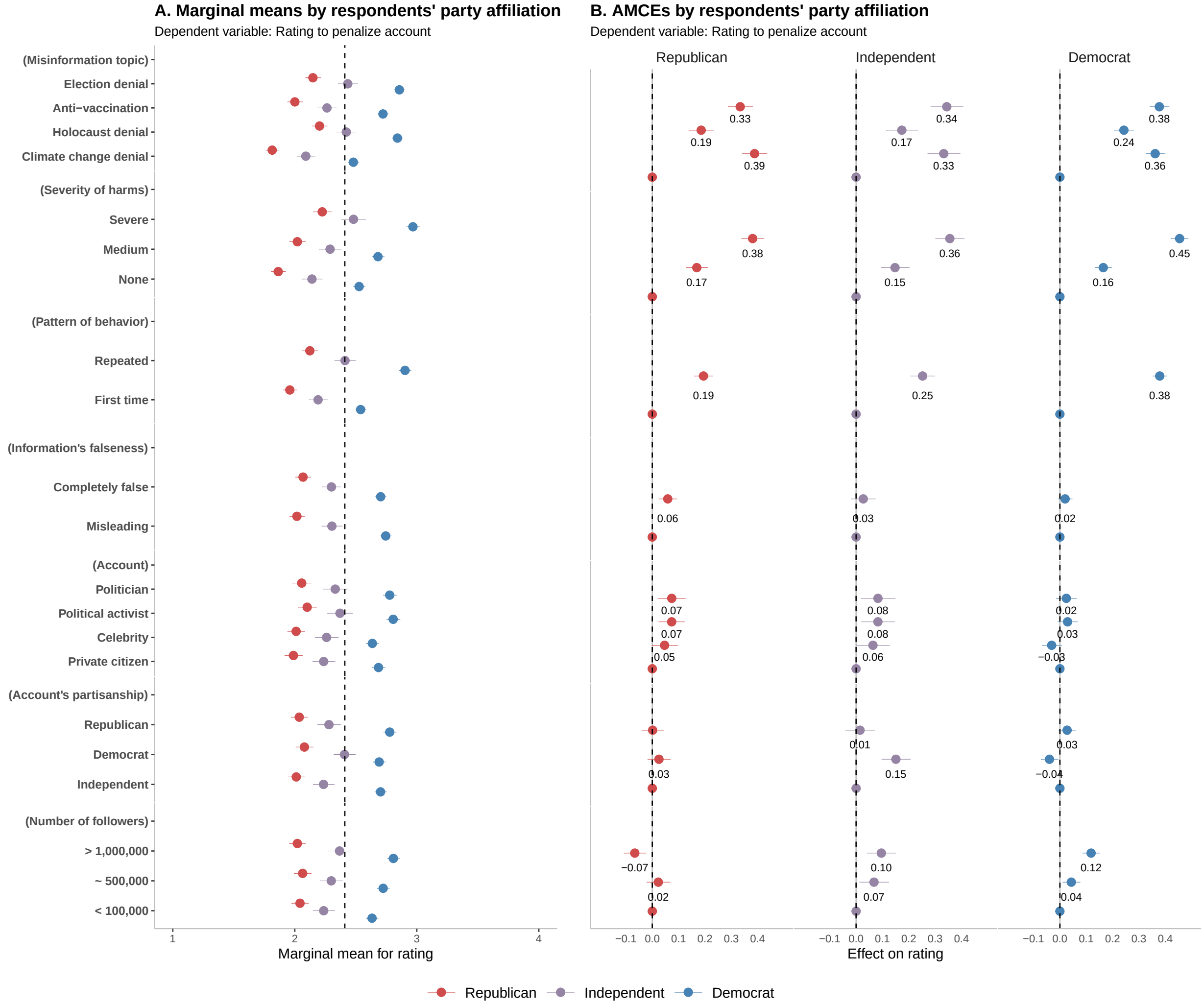


Figure A4: *Respondent subgroup analyses: Rating by respondents' party affiliation.* Marginal means point estimates and average marginal component effects (AMCEs) plotted with 95% CIs. Panel A: Marginal means represent the average rating for decisions to penalize the account for each attribute level for three respondent subgroups: Republicans, Independents, and Democrats. Dashed line represents the grand mean for rating (2.41). Panel B: AMCEs represent effects on rating to penalize the account for each attribute level, faceted by three subgroups: Republicans, Independents, and Democrats. Dashed lines represent the null effect.

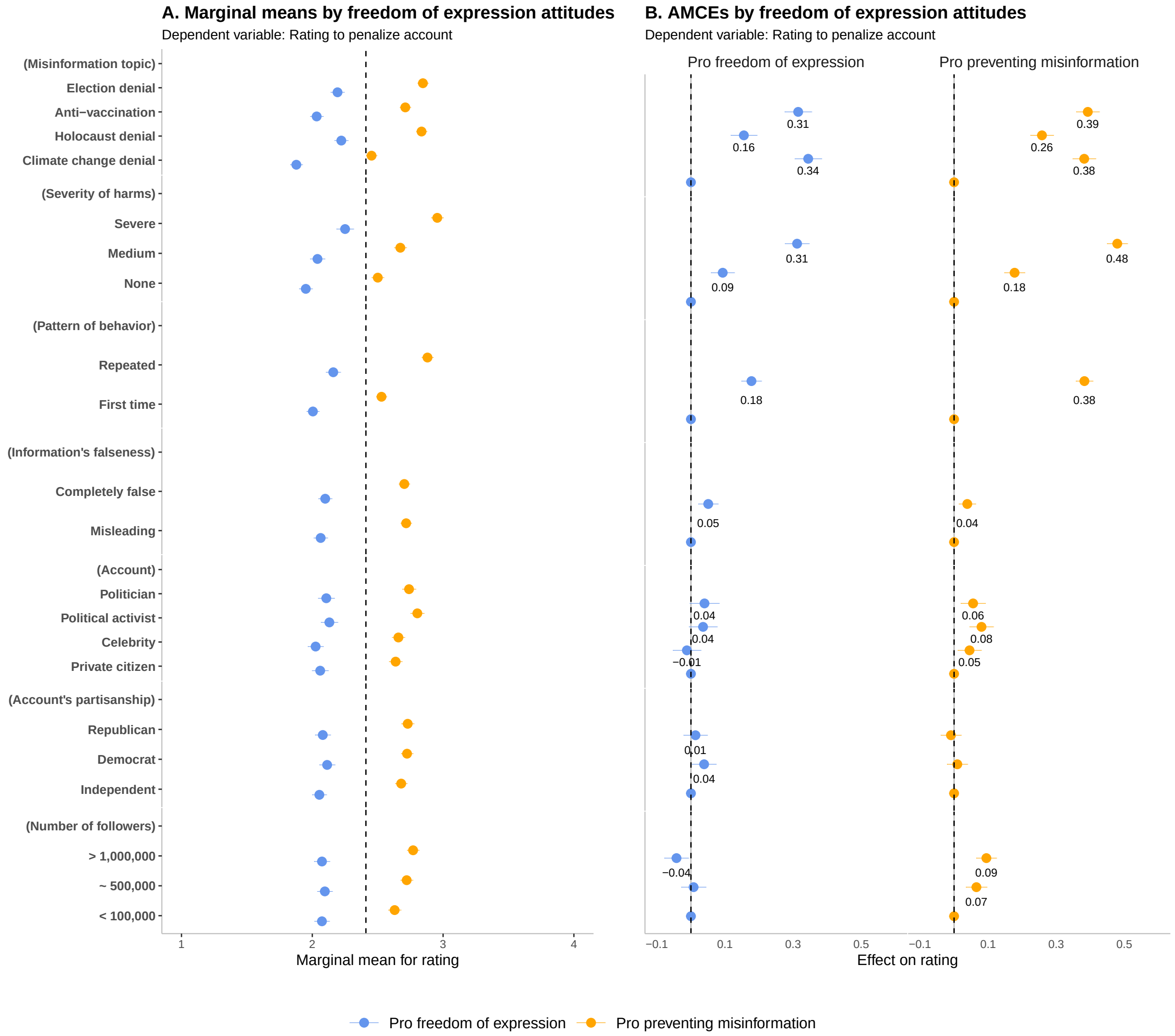


Figure A5: *Respondent subgroup analyses: Rating by respondents' attitudes toward free speech.* Marginal means point estimates and average marginal component effects (AMCEs) plotted with 95% CIs. Panel A: Marginal means represent the average rating for decisions to penalize the account for each attribute level for two respondent subgroups: pro-freedom of expression and pro-mitigating misinformation. Dashed line represents the grand mean for rating (2.41). Panel B: AMCEs represent effects on rating to penalize the accounts for each attribute level, faceted by two respondent subgroups: Pro-freedom of expression and pro-mitigating misinformation. Dashed lines represent the null effect.

Appendix B

Descriptive and Summary Statistics on Survey Measures

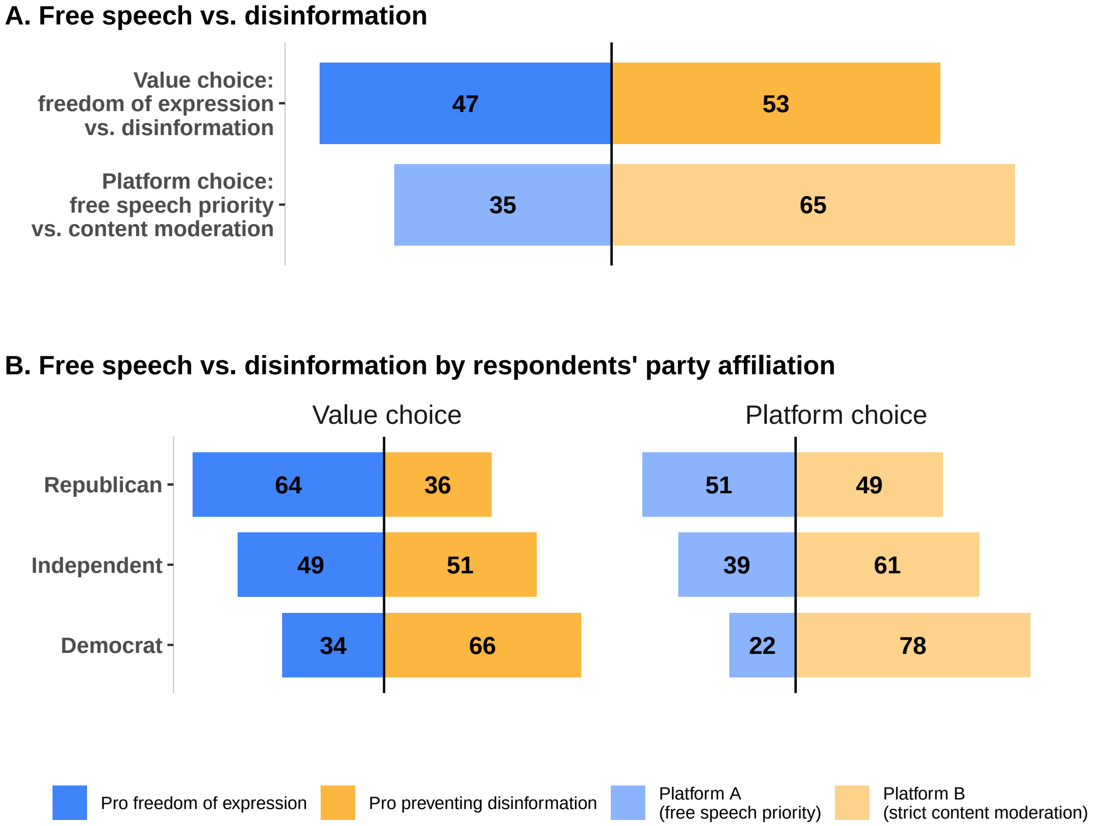


Figure B1: *Preferences on freedom of expression for values and platforms.* Value choice: "If you absolutely have to choose between protecting freedom of expression and preventing disinformation from spreading, which is more important to you?" Platform choice: "Imagine you are considering joining one of two rival social media platforms. Platform A claims that it will always prioritize free speech and will never suspend an account or remove a post that incites violence, constitutes hate speech, or spreads false information. Platform B has a zero tolerance policy against false information, hate speech, and incitement to violence, and it will enforce strict content moderation rules for everyone. Which social media platform would you rather join?" Panel A: Proportion of responses for both questions for all participants. Panel B: Proportions by respondents' party affiliation.

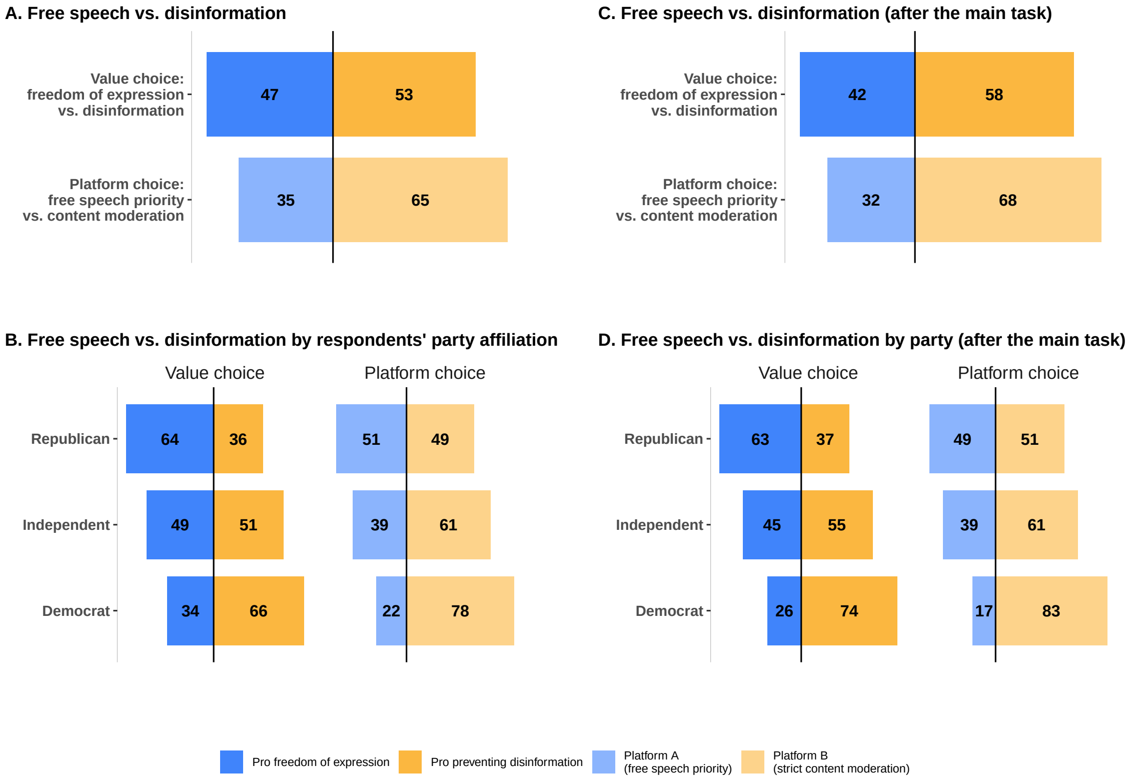
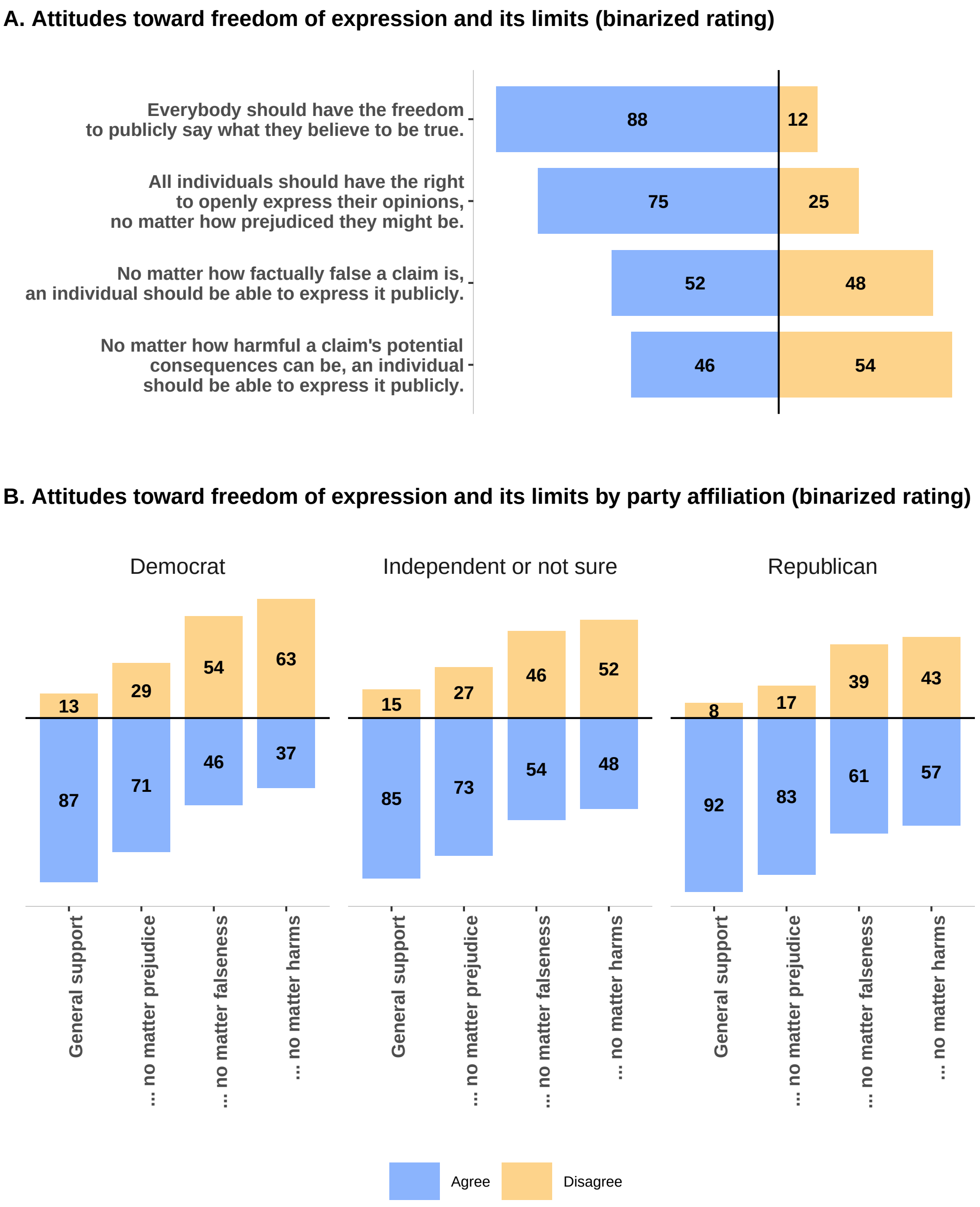


Figure B2: *Freedom of expression versus mitigating misinformation: Before and after main task.* All numeric values represent percentages. Value choice: “If you absolutely have to choose between protecting freedom of expression and preventing disinformation from spreading, which is more important to you?”. Platform choice: “Imagine you are considering joining one of two rival social media platforms. Platform A claims that it will always prioritize free speech and will never suspend an account or remove a post that incites violence, constitutes hate speech, or spreads false information. Platform B has a zero tolerance policy against false information, hate speech, and incitement to violence, and it will enforce strict content moderation rules for everyone. Which social media platform would you rather join?” Panel A: Proportion of responses for both questions for all participants, before the main study task. Panel B: Proportions by party affiliation, before the main study task. Panel C: Proportion of responses for both questions for all participants, after the main study task. Panel D: Proportions by respondents’ party affiliation, after the main study task.



A. Attitudes toward freedom of expression and its limits (binarized rating)

B. Attitudes toward freedom of expression and its limits by party affiliation (binarized rating)

Figure B3: *Freedom of expression and its limits.* All numeric values represent percentages. The four items addressed participants' general attitudes toward freedom of expression and its limits in cases of prejudice, falsehoods, and potential for harm on a 6-point Likert scale (Strongly disagree, Moderately disagree, Slightly disagree, Slightly agree, Moderately agree, Strongly agree). In this figure, these responses are grouped in two categories: Agree and Disagree. Panel A: Proportions of responses to four items querying general attitudes toward freedom of expression and its limits. Panel B: Proportions by respondents' party affiliation.

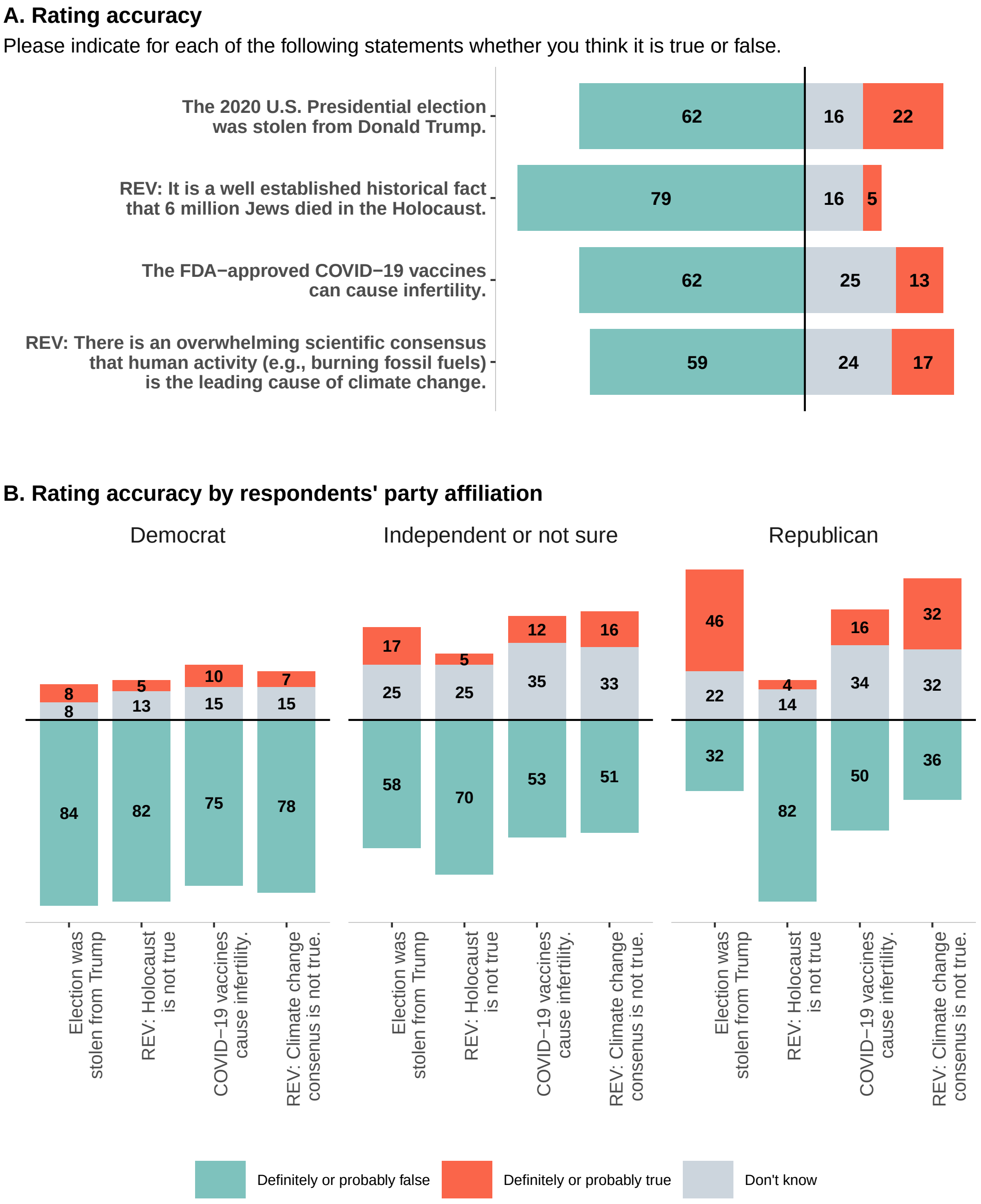


Figure B4: *Accuracy ratings for misinformation statements.* All numeric values represent percentages. Panel A: Proportions of responses for rating accuracy of four claims on a 5-point Likert scale (definitely false, probably false, don't know, probably true, definitely true). Responses are grouped into three categories: Definitely or probably false: Do not know; Definitely or probably true. Responses for accurate statements are reverse coded (denoted by "REV" before the statement). Panel B: Proportions by respondents' party affiliation.

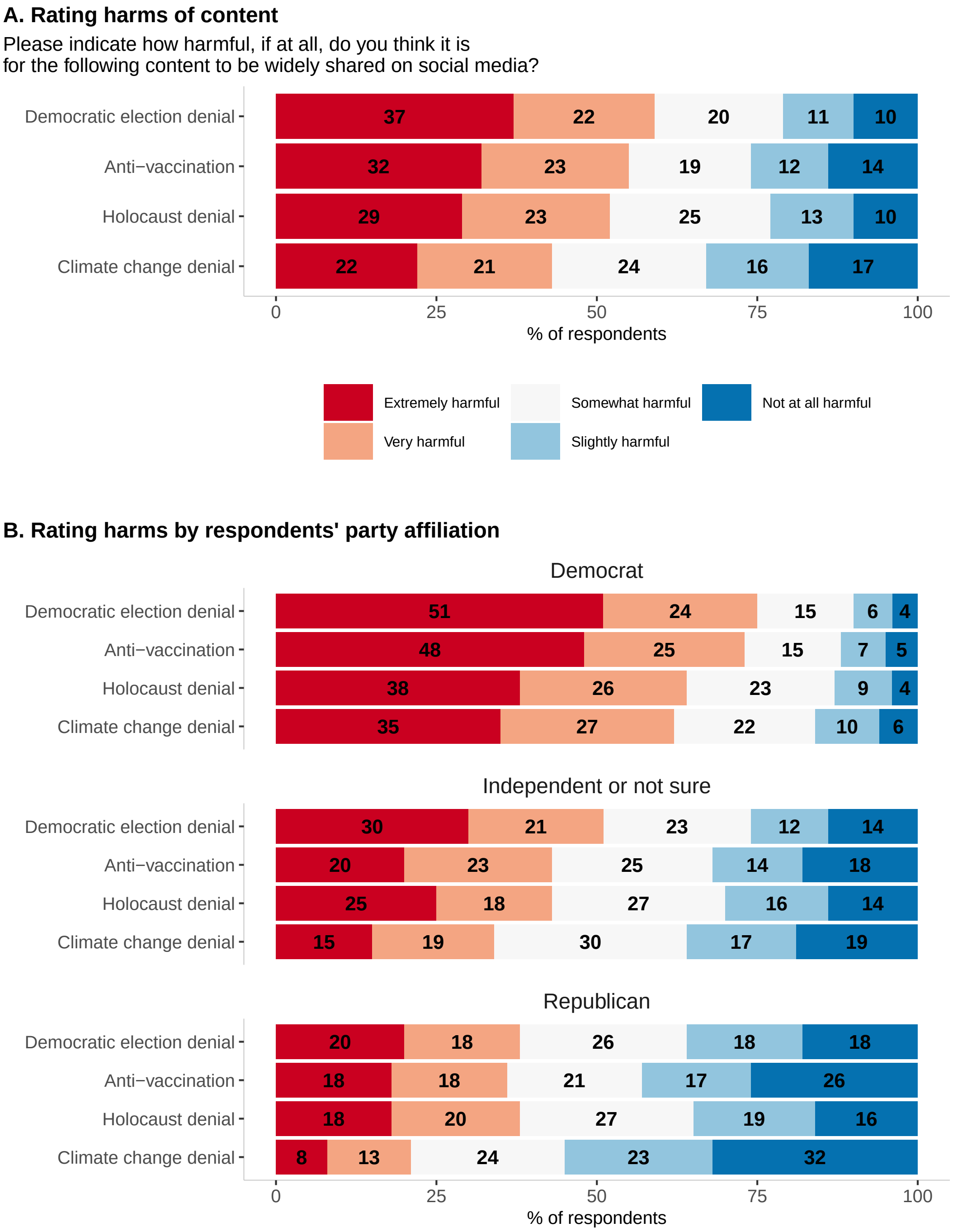


Figure B5: *Content harm ratings for statements relevant to scenarios.* All numeric values represent percentages. Panel A: Proportions of responses for rating of perceived harm of the content featured in each scenario on a 5-point Likert scale (not at all harmful, slightly harmful, somewhat harmful, very harmful, extremely harmful). Panel B: Proportions by respondents' party affiliation.

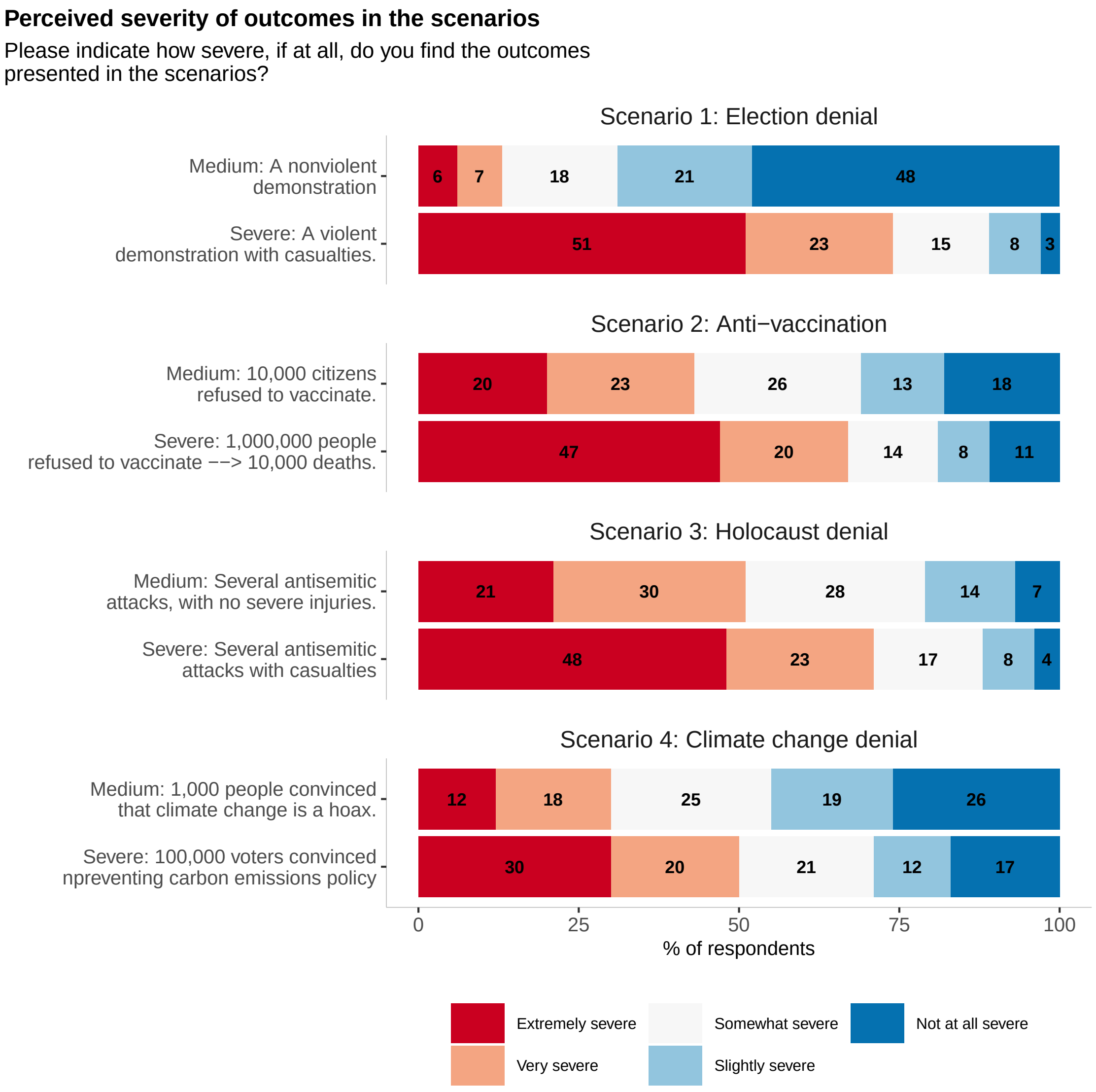


Figure B6: *Perceived severity of outcomes in the scenarios.* All numeric values represent percentages. Proportions of responses for rating of perceived severity of the outcomes featured in the scenarios (on a 5-point Likert scale).

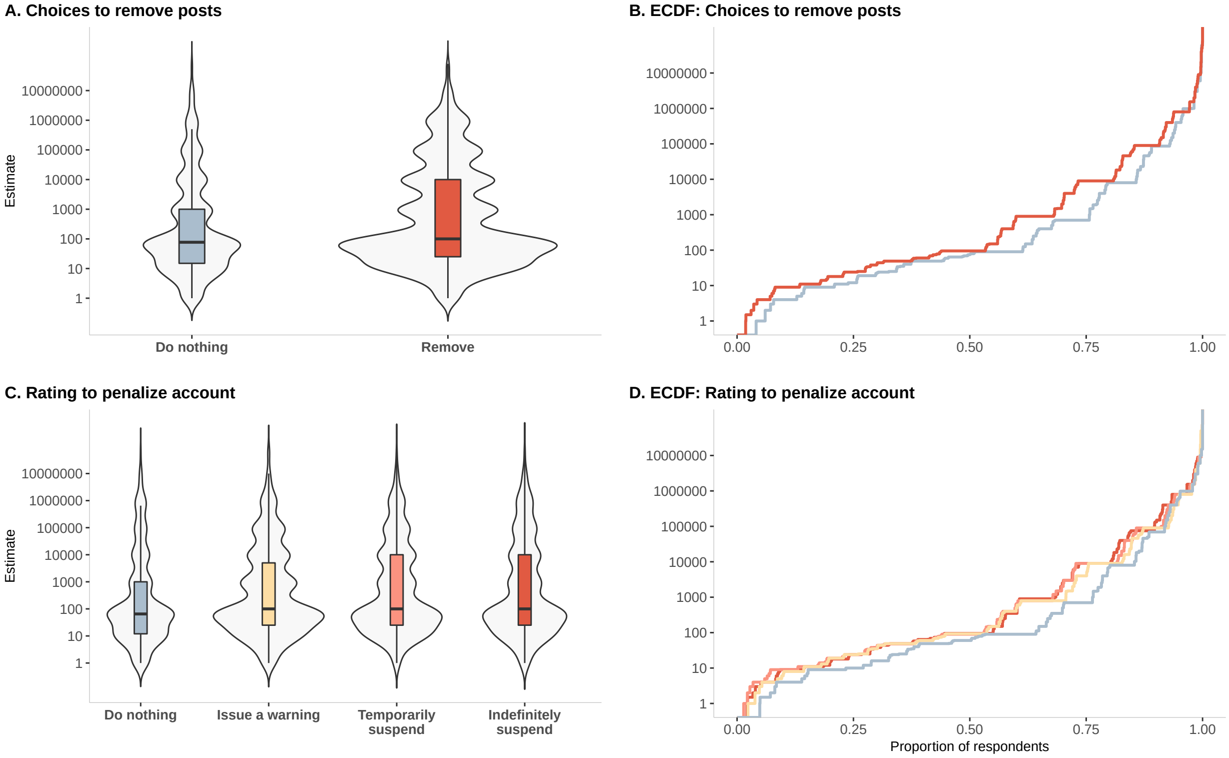


Figure B7: *Summary statistics for relationship between estimates of number of disinformation accounts and outcome measures in the scenarios.* The estimation question was: “To the best of your knowledge, how many individuals are responsible for 65% of the anti-vaccination disinformation on Facebook and Twitter? Please indicate or estimate a number.” Panel A: Boxplots and violin plots for the choice to remove posts. Panel C: Boxplots and violin plots for the choice to penalize account. Panel B: Empirical cumulative density function (ECDF) for the choice to remove posts. Panel D: Empirical cumulative density function (ECDF) for rating to penalize account.

Appendix C

Subset Analyses Based on the Accuracy of Knowledge of the Facts at Hand

For the robustness analyses, we partitioned our conjoint data depending on whether a respondent's belief about the facts at hand was accurate or not. For instance, in the anti-vaccination scenario, which was displayed in 10,256 cases with different combinations of attributes, we excluded all conjoint data where a respondent indicated in the post-treatment question that the statement "The FDA-approved COVID-19 vaccines can cause infertility" was definitely or possibly true or if they did not know whether it was true or false (thus retaining 6,368 random cases of this scenario). For the election denial scenario, we excluded profiles where respondents indicated in the post-treatment measure that the statement "The 2020 U.S. Presidential election was stolen from Donald Trump" was definitely or possibly true or they did not know. In the Holocaust denial scenario, we excluded profiles where respondents indicated that the statement "It is a well established historical fact that 6 million Jews died in the Holocaust" was definitely or probably false or if they did not know. In the climate change denial scenario, we excluded profiles where respondents indicated that the statement "There is an overwhelming scientific consensus that human activity (e.g., burning fossil fuels) is the leading cause of climate change" was definitely or probably false or they did not know. From the total set of 40,845 cases we thus created a subset of 26,726 profiles with accurate matching beliefs.

As Appendix Figure C1 shows, more respondents in the subset than in the full dataset (Figure 2) opted to remove false and misleading posts and to penalize accounts that spread them. In the subset, the election denial scenario and the anti-vaccination scenario overtook the Holocaust denial scenario in participants' likelihood to take action. Note, however, that these changes are difficult to interpret because the subset analysis excluded responses whenever a respondent endorsed inaccurate beliefs and the accuracy of beliefs systematically differed between the scenarios, the partisan groups, and their interaction. The majority of Republicans with accurate beliefs were more likely to sanction online misinformation than were the Republicans in the full dataset. However, the main patterns in the subgroup differences remained (Appendix Figures C3 and C4), including the finding that Republicans were less likely than Independents and Democrats to take action against misinformation.

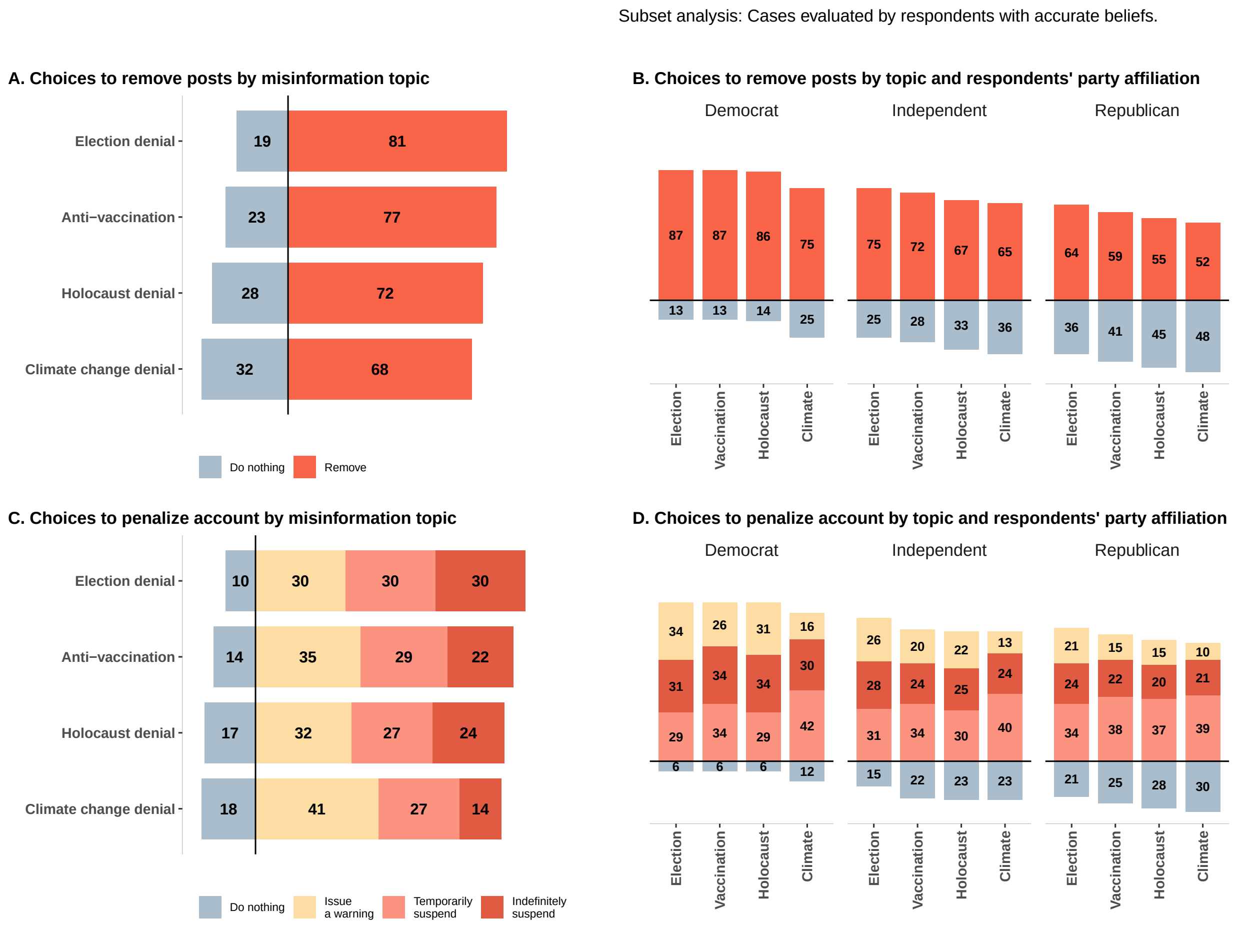


Figure C1: *Subset of cases evaluated by respondents with accurate beliefs: Proportions.* All numeric values represent percentages. Panel A: Choices to remove the posts or do nothing by misinformation topic. Panel B: Choices to remove the posts or do nothing, by topic and party affiliation. Panel C: Choices to penalize the account by misinformation topic. Panel D: Choices to penalize the account, by topic and party affiliation. Total N of cases in the subset: 26,726 (evaluated by Democrats: 15,351; by Independents: 4,769; by Republicans: 6,606).

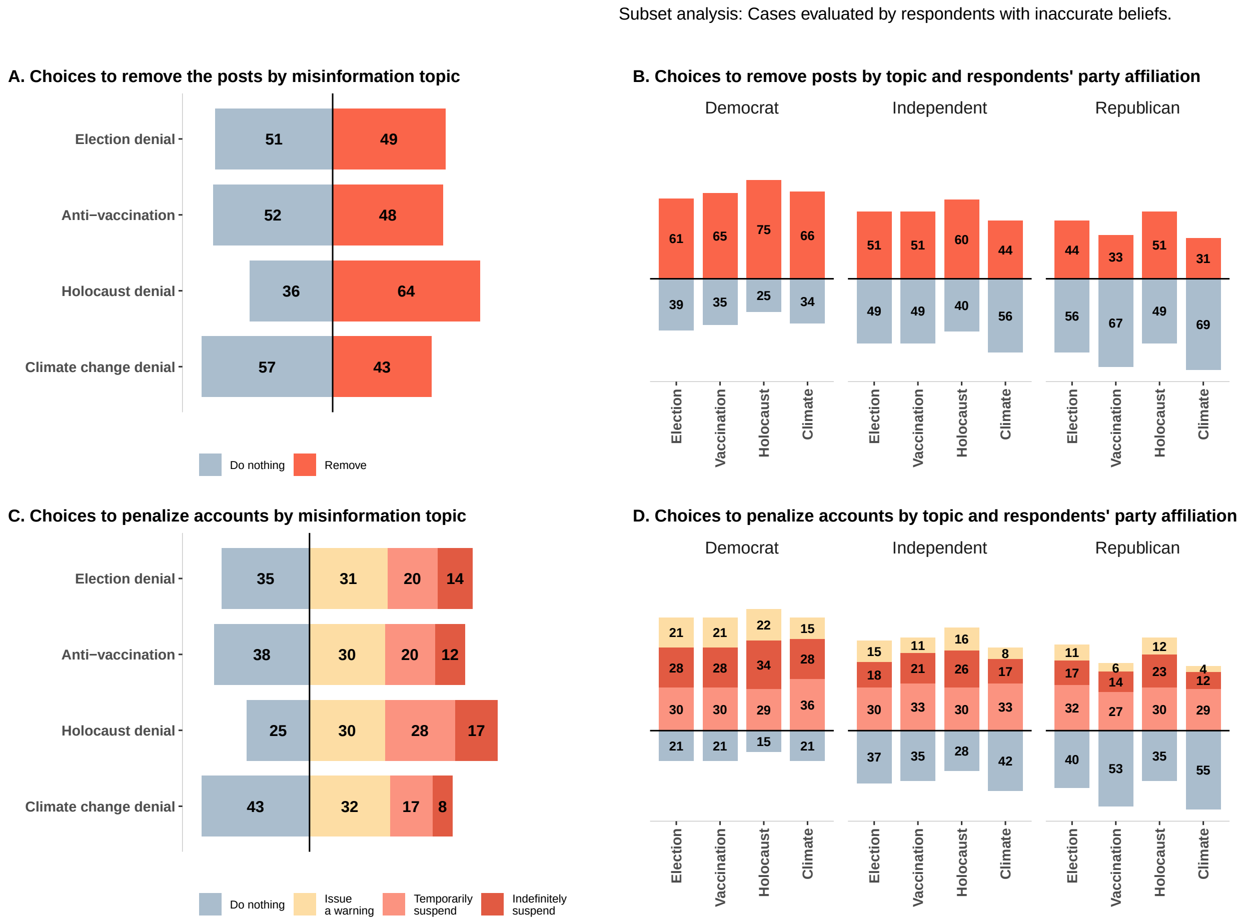


Figure C2: *Subset of cases evaluated by respondents with inaccurate beliefs: Proportions.* All numeric values represent percentages. Panel A. Choices to remove the posts or do nothing by misinformation topic. Panel B: Choices to remove the posts or do nothing, by topic and party affiliation. Panel C: Choices to penalize the account by misinformation topic. Panel D: Choices to penalize the account, by topic and party affiliation. Total N of cases in the subset: 14,119 (evaluated by Democrats: 3,987; by Independents: 3,460; by Republicans: 6,672).

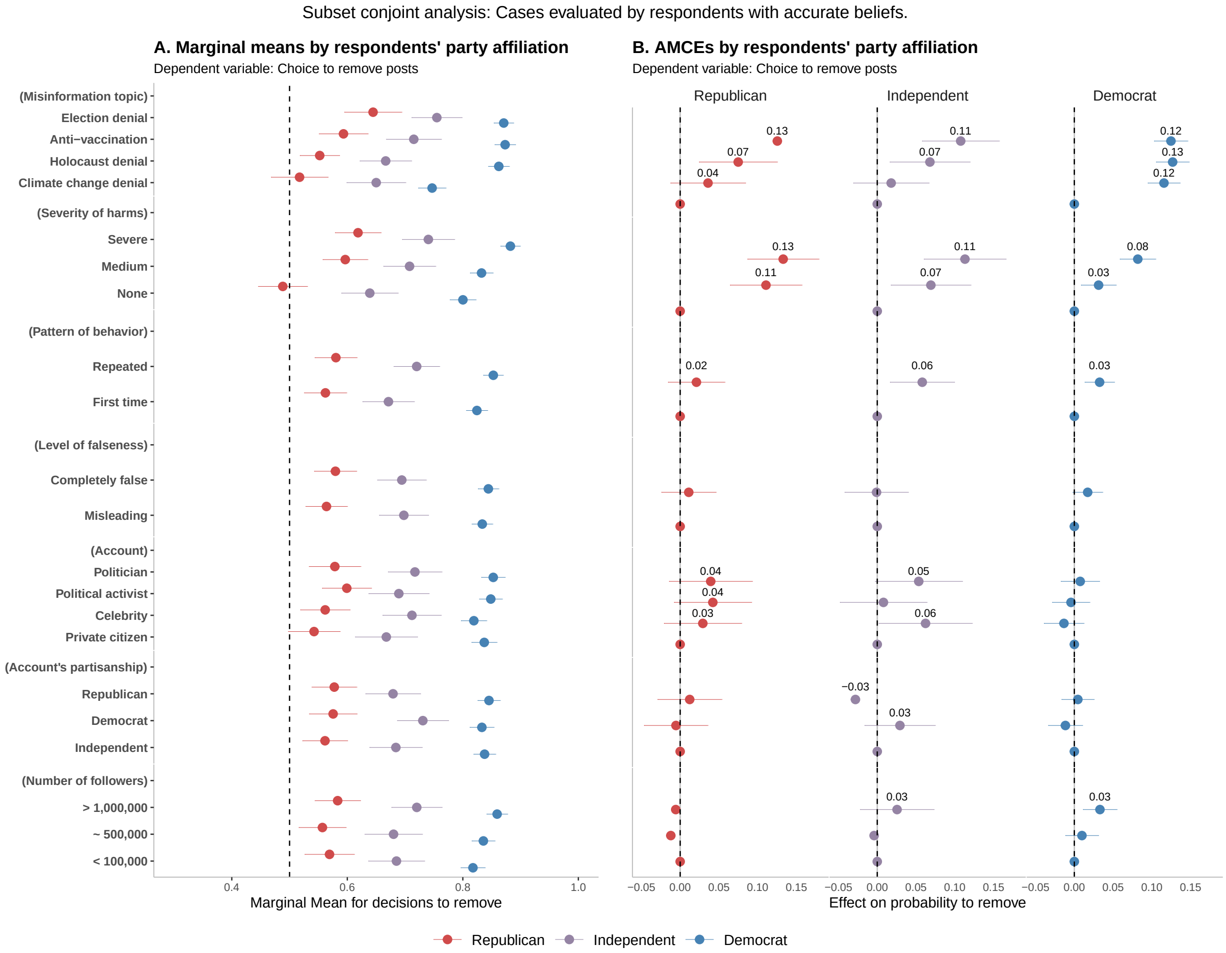


Figure C3: *Subset of profiles evaluated by respondents with accurate beliefs: Subgroup conjoint analysis by respondents' party affiliation (choice).* Marginal means point estimates and average marginal component effects (AMCEs) plotted with 95% CIs. Panel A: Marginal means represent the average likelihood of decisions to remove the posts for each attribute level for three respondent subgroups: Republicans, Independents, and Democrats. Dashed line represents the mean value for a binary decision (0.5). Panel B: AMCEs represent effects on probability to remove the posts for each attribute level, faceted by three respondent subgroups: Republicans, Independents, and Democrats. Dashed lines represent the null effect. Only estimates larger than ±0.02 are labeled. Total N of cases in the subset = 26,726 (evaluated by Democrats: 15,325; by Independents: 4,722; by Republicans: 6,652).

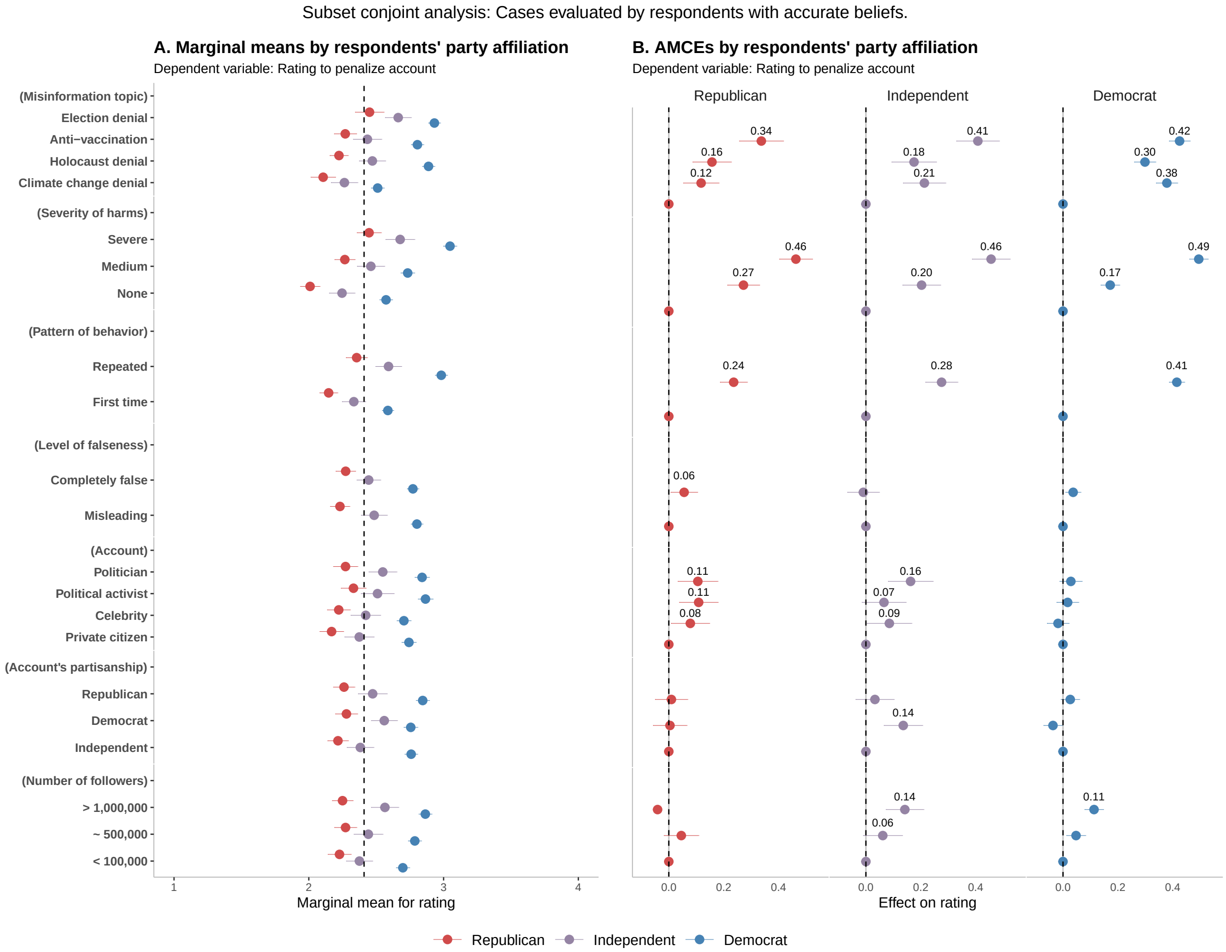


Figure C4: *Subset of profiles evaluated by respondents with inaccurate beliefs: Subgroup conjoint analysis by respondents' party affiliation (rating).* Marginal means point estimates and average marginal component effects (AMCEs) plotted with 95% CIs. Panel A: Marginal means represent the average rating for decisions to penalize the account for each attribute level for three respondent subgroups: Republicans, Independents, and Democrats. Dashed line represents the grand mean for rating (2.41). Panel B: AMCEs represent effects on rating to penalize the account for each attribute level, faceted by three respondent subgroups: Republicans, Independents, and Democrats. Dashed lines represent the null effect. Only estimates larger than ±0.05 are labeled. Total N of cases in the subset = 26,726 (evaluated by Democrats: 15,325; by Independents: 4,722; by Republicans: 6,652).

Appendix D
Misinformation Policies

Table D1

Social Media Platforms' Misinformation Policies (last updated June 03, 2022)

Type of content	Platform	Policy	Strike system	Cases
COVID-19 misinformation	Google: Ads	Misrepresentation policy: What is not allowed: "Content promoting harmful health claims, or content that relates to a current, major health crisis and contradicts authoritative scientific consensus. Examples (non-exhaustive): Anti-vaccine advocacy; denial of the existence of medical conditions such as AIDS or Covid-19; gay conversion therapy"(Google, n.d.-d)	"Violations of this policy will not lead to immediate account suspension without prior warning." A warning will be issued at least 7 days prior to suspension and appeal is possible." (Google, n.d.-d)	
	Google: YouTube	COVID-19 medical misinformation policy: "YouTube doesn't allow content that spreads medical misinformation that contradicts local health authorities' (LHA) or the World Health Organization's (WHO) medical information about COVID-19." (Google, n.d.-b) Vaccine misinformation policy: "YouTube doesn't allow content that poses a serious risk of egregious harm by spreading medical misinformation about currently administered vaccines that are approved and confirmed to be safe and effective by local health authorities and by the World Health Organization (WHO)." (Google, n.d.-e)	Post deletion. 1st offense: No penalty, warning 2nd offense, 1 strike: 1 week ban on activity 2 strikes in 90 days: 2 week ban on posting 3 strikes in 90 days: Channel termination * (Google, n.d.-a)	"Since last year, we've removed over 130,000 videos for violating our COVID-19 vaccine policies."* (The YouTube Team, 2021)

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
	Meta: Ads	Misleading Claims Advertising Policy: Prohibits ads that "make deceptive, false or unsubstantiated health claims, including claims that a product or service can provide 100% prevention or immunity, or is a cure for the virus." (Meta, n.d.-a)	"Repeat offenders are subject to enforcement. If we see advertisers repeatedly violate our advertising policies, we may take action, including but not limited to, losing the ability to advertise via disablement of a single ad account, Ads Manager, Business Manager, Facebook Page or Instagram page." (Meta, n.d.-a)	"Today, following consultations with leading health organizations, including the World Health Organization (WHO), we are expanding the list of false claims we will remove to include additional debunked claims about the coronavirus and vaccines. This includes claims such as: COVID-19 is man-made or manufactured; Vaccines are not effective at preventing the disease they are meant to protect against; It's safer to get the disease than to get the vaccine; Vaccines are toxic, dangerous or cause autism, The full list of claims is available here, and we already prohibit these claims in ads. These new policies will help us continue to take aggressive action against misinformation about COVID-19 and vaccines." (Rosen, 2020)

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
	Meta: Facebook, Instagram	Restricted Goods and Services policy: Prohibits posts that "indicate a sense of urgency or claims that prevention is guaranteed." (Instagram, n.d.) Hate speech policy: Prohibits posts that "state that people who share a protected characteristic such as race or religion have the virus, created the virus or are spreading the virus." (Instagram, n.d.) Bullying and harassment policy: Prohibits "claims that a private individual has COVID-19, unless that person has self-declared or information about their health status is publicly available." (Instagram, n.d.) List of measures against COVID-19 misinformation: "We remove COVID-19 related misinformation that could contribute to imminent physical harm." (Clegg, 2020)	Posts violating community guidelines get deleted. One strike: Warning and no further restrictions. 2 strikes: One-day restriction from creating content, such as posting, commenting, using Facebook Live or creating a Page. 3 strikes: 3-day restriction from creating content. 4 strikes: 7-day restriction from creating content. 5 or more strikes: 30-day restriction from creating content. All strikes on Facebook or Instagram expire after one year.** (Meta, n.d.-c)	"During the month of March, we displayed warnings on about 40 million posts related to COVID-19 on Facebook, based on around 4,000 articles by our independent fact-checking partners. When people saw those warning labels, 95% of the time they did not go on to view the original content. To date, we've also removed hundreds of thousands of pieces of misinformation that could lead to imminent physical harm. Examples of misinformation we've removed include harmful claims like drinking bleach cures the virus and theories like physical distancing is ineffective in preventing the disease from spreading." (Rosen, 2020)
	Twitter	COVID-19 misleading information policy: "Content that is demonstrably false or misleading and may lead to significant risk of harm (such as increased exposure to the virus, or adverse effects on public health systems) may not be shared on Twitter." (Twitter, n.d.-b)	Labeling (1 strike), Request for Tweet deletion (2 strikes). 1 strike: No account-level action 2 strikes: 12-hour account lock 3 strikes: 12-hour account lock 4 strikes: 7-day account lock 5 or more strikes: Permanent suspension *** (Twitter, n.d.-b)	"Since introducing these policies on March 18, we have removed more than 1,100 Tweets containing misleading and potentially harmful content from Twitter. Additionally, our automated systems have challenged more than 1.5 million accounts which were targeting discussions around COVID-19 with spammy or manipulative behaviors." (Twitter, 2020)

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
	TikTok	Community guidelines - COVID-19: “Misinformation is defined as content that is inaccurate or false. While we encourage our community to have respectful conversations about subjects that matter to them, we do not permit misinformation that causes harm to individuals, our community, or the larger public regardless of intent. Do not post, upload, stream, or share ... medical misinformation that can cause harm to an individual’s physical health” (TikTok, n.d.-b)	1st violation: Warning; if the violation is under zero-tolerance policy, then automatic ban + may also block a device to help prevent future accounts from being created. 2nd violation: One or more of the following; • Temporary ban (typically between 24 or 48 hours), depending on the severity of the violation and previous violations. • Restrict the account to a view-only experience (typically between 72 hours or up to one week) • Permanent ban But: Accrued violations will expire from individuals’ record over time **** (TikTok, n.d.-a)	“We removed 51,505 videos in the second half of 2020 for promoting COVID-19 misinformation. Of those videos, 86% were removed before they were reported to us, 87% were removed within 24 hours of being uploaded to TikTok, and 71% had zero views.” (TikTok, 2021)
	Spotify	Spotify Platform Rules: “What to avoid: ... Content that promotes dangerous false or dangerous deceptive medical information that may cause offline harm or poses a direct threat to public health.” (Spotify, 2022)	“Breaking the rules may result in the violative content being removed from Spotify. Repeated or egregious violations may result in accounts being suspended and/or terminated.” (Spotify, 2022)	
	Pinterest	Community gudelines - Misinformation: “We remove or limit distribution of false or misleading content that may harm Pinners’ or the public’s well-being, safety or trust, including: Medically unsupported health claims that risk public health and safety, including the promotion of false cures, anti-vaccination advice, or misinformation about public health or safety emergencies” (Pinterest, n.d.-b)	“We make sure content meets our Community Guidelines through both automated processes and human review. Accounts may be suspended due to single or repeat violations of our Community Guidelines”***** (Pinterest, n.d.-a)	

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
Democratic election denial and misinformation on the voting process	Google: Ads	Misrepresentation policy: “The following is not allowed: Making claims that are demonstrably false and could significantly undermine participation or trust in an electoral or democratic process. Example (non-exhaustive): Information about public voting procedures, political candidate eligibility based on age or birthplace, election results, or census participation that contradicts official government records; incorrect claims that a public figure has died, or been involved in an accident.” (Google, n.d.-d)	“Violations of this policy will not lead to immediate account suspension without prior warning.” A warning will be issued at least 7 days prior to suspension and appeal is possible.” (Google, n.d.-d)	“After review, and in light of concerns about the ongoing potential for violence, we removed new content uploaded to Donald J. Trump’s channel for violating our policies. It now has its 1st strike & is temporarily prevented from uploading new content for a *minimum* of 7 days.” (YouTube Insider, 2021)
	Google: YouTube	Community guidelines: The content removed may include “- Content that aims to mislead people about voting or the census processes, like telling viewers an incorrect voting date. - Content that advances false claims related to the technical eligibility requirements for current political candidates and sitting elected officials to serve in office, such as false claims that a candidate is not eligible to hold office based on false information about citizenship status requirements to hold office in that country. - Content that advances false claims that widespread fraud, errors, or glitches changed the outcome of any past U.S. presidential election.” (YouTube, n.d.)	* (Google, n.d.-a)	
	Meta: Facebook, Instagram	Coordinating harm and publicising crime: “In an effort to prevent and disrupt offline harm and copycat behaviour, we prohibit people from facilitating, organising, promoting or admitting to certain criminal or harmful activities targeted at people, businesses, property or animal. ... Do not post content that falls into the following categories: ... - Voter and/or census interference” (Meta, n.d.-b)	** (Meta, n.d.-c) “When there is civil unrest, we may also restrict accounts by public figures for longer periods of time when they incite or praise ongoing violence. We’ll determine the restriction period after assessing the severity of the violation, the account’s history of past violations and the overall risk to public safety.” (Meta, n.d.-c)	“Given the gravity of the circumstances that led to Mr. Trump’s suspension, we believe his actions constituted a severe violation of our rules which merit the highest penalty available under the new enforcement protocols. We are suspending his accounts for two years, effective from the date of the initial suspension on January 7 this year.” (Clegg, 2021a)

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
	Twitter	Civic integrity policy: “You may not use Twitter’s services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process. In addition, we may label and reduce the visibility of Tweets containing false or misleading information about civic processes in order to provide additional context.” (Twitter, n.d.-a)	*** (Twitter, n.d.-a)	“After close review of recent Tweets from the @realDonaldTrump account and the context around them — specifically how they are being received and interpreted on and off Twitter — we have permanently suspended the account. ... President Trump’s statement that he will not be attending the Inauguration is being received by a number of his supporters as further confirmation that the election was not legitimate and is seen as him disavowing his previous claim made via two Tweets (1, 2) by his Deputy Chief of Staff, Dan Scavino, that there would be an “orderly transition” on January 20th.” (Twitter, 2021)
	TikTok	Community guidelines - Election integrity: “Misinformation is defined as content that is inaccurate or false. While we encourage our community to have respectful conversations about subjects that matter to them, we do not permit misinformation that causes harm to individuals, our community, or the larger public regardless of intent. Do not post, upload, stream, or share: ... - Content that misleads community members about elections or other civic processes ... - Misinformation related to emergencies that induces panic.” (TikTok, 2020b)	**** (TikTok, n.d.-a)	“In the second half of 2020, 347,225 videos were removed in the US for election misinformation, disinformation, or manipulated media. We worked with fact checkers at PolitiFact, Lead Stories, and SciVerify to assess the accuracy of content and limit distribution of unsubstantiated content. As a result, 441,028 videos were not eligible for recommendation into anyone’s For You feed. We further removed 1,750,000 accounts that were used for automation during the timeframe of the US elections.” (TikTok, 2021)
	Spotify	Spotify Platform Rules: “Content that attempts to manipulate or interfere with election-related processes includes, but may not be limited to: - misrepresentation of procedures in a civic process that could discourage or prevent participation - misleading content promoted to intimidate or suppress voters from participating in an election” (Spotify, 2022)	“Breaking the rules may result in the violative content being removed from Spotify. Repeated or egregious violations may result in accounts being suspended and/or terminated.” (Spotify, 2022)	

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
	Pinterest	Community gudelines - Civic participation misinformation: “We remove or limit distribution of false or misleading content that may harm Pinners’ or the public’s well-being, safety or trust, including: False or misleading content that impedes an election’s integrity or an individual’s or group’s civic participation ... - about who can vote or participate in the census and what information must be provided to participate” (Pinterest, n.d.-b)	***** (Pinterest, n.d.-a)	
Holocaust denial	Google: YouTube	Hate speech policy: “Don’t post content on YouTube if the purpose of that content is to do one or more of the following: ... - Deny that a well-documented, violent event took place.” (Google, n.d.-c)	* (Google, n.d.-a) And additionally: “If we think your content comes close to hate speech, we may limit YouTube features available for that content” (Google, n.d.-c) with no: comments, suggested videos, likes.	In an interview with the NPR national security correspondent Hannah Allam: “Well, there was no waiting around. This policy kicked in immediately. YouTube videos with extremist content started vanishing - videos that promoted white supremacy, neo-Nazi videos. Some civil rights groups and people who’ve been targeted for harassment online say it’s a step in the right direction, although they also have concerns that it doesn’t go far enough or it’s impossible to enforce. And on the flipside, there are people who say it goes too far.” (Garcia-Navarro, 2019)
	Meta: Facebook, Instagram	Hate speech policy: Do not post “Designated dehumanising comparisons, generalisations or behavioural statements (in written or visual form) that include: ... - Denying or distorting information about the Holocaust.” (Meta, n.d.-d)	** (Meta, n.d.-b)	

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
	Twitter	Abusive behavior policy: “We prohibit content that denies that mass murder or other mass casualty events took place, where we can verify that the event occured, and when the content is shared with abusive intent. This may include references to such an event as a “hoax” or claims that victims or survivors are fake or “actors.” It includes, but is not limited to, events like the Holocaust, school shootings, terrorist attacks, and natural disasters.” (Twitter, n.d.)	“When determining the penalty for violating this policy, we consider a number of factors including, but not limited to the severity of the violation and an individual’s previous record of rule violations. The following is a list of potential enforcement options for content that violates this policy: - Downranking Tweets in replies, except when the user follows the Tweet author. - Making Tweets ineligible for amplification in Top search results and/or on timelines for users who don’t follow the Tweet author. - Excluding Tweets and/or accounts in email or in-product recommendations. - Requiring Tweet removal. - Suspending accounts.” (Twitter, n.d.)	
	TikTok	Community Guidelines - Hateful behavior: “Do not post, upload, stream, or share: ... Content that denies well-documented and violent events have taken place affecting groups with protected attributes.” (TikTok, 2020a)	**** (TikTok, n.d.-a)	
	Spotify	No policy yet.		
	Pinterest	Community gudelines - Hateful activities: “We limit the distribution of or remove such content and accounts, including: Hate-based conspiracy theories and misinformation, like Holocaust denial” (Pinterest, n.d.-b)	***** (Pinterest, n.d.-a)	

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
Climate change denial	Google: Ads	Misrepresentation policy: “We want users to trust the ads on our platform, so we strive to ensure ads are clear and honest, and provide the information that users need to make informed decisions. We don’t allow ads or destinations that deceive users by excluding relevant product information or providing misleading information about products, services, or businesses [e.g.,] - Making claims that contradict authoritative, scientific consensus on climate change” (Google, n.d.-d)	- Ad or extension disapproval until the issue is resolved - Account suspension with (notification will be sent at least 7 days prior to suspension action) or without warning (if and only if egregious violation of the Google Ads policies happens) - Remarketing list disabling - Compliance review of the profile (Google, n.d.)	
	Google: YouTube	No policy yet.		
	Meta: Facebook, Instagram	No policy yet.		“We have a responsibility to tackle climate misinformation on our services, which is why we partner with more than 80 independent fact-checking organizations globally to review and rate content, including content about climate change. When they rate content as false, we reduce its distribution so fewer people see it and we show a warning label with more context. And we apply penalties to people who repeatedly share false information.” (Clegg, 2021b)
	Twitter	No policy yet.		
	TikTok	No policy yet.		
	Spotify	No policy yet.		

Continued on next page

Table D1 – continued from previous page

Type of content	Platform	Policies	System of removal	Cases
	Pinterest	Community gudelines - Climate misinformation: “We remove or limit distribution of false or misleading content that may harm Pinners’ or the public’s well-being, safety or trust, including: - Content that denies the existence or impacts of climate change, the human influence on climate change, or that climate change is backed by scientific consensus. - False or misleading content about climate change solutions that contradict well-established scientific consensus. - Content that misrepresents scientific data, including by omission or cherry-picking, in order to erode trust in climate science and experts.” (Pinterest, n.d.-b)	***** (Pinterest, n.d.-a)	

References

Clegg, N. (2020, March 25). *Combating COVID-19 misinformation across our apps.* Meta. https://about.fb.com/news/2020/03/combating-covid-19-misinformation/

Clegg, N. (2021a, June 4). *In response to Oversight Board, Trump suspended for two years; will only be reinstated if conditions permit.* Meta. https://about.fb.com/news/2021/06/facebook-response-to-oversight-board-recommendations-trump/

Clegg, N. (2021b, November 1). *Our commitment to combating climate change.* Meta. https://about.fb.com/news/2021/11/our-commitment-to-combating-climate-change/

Garcia-Navarro, L. (2019, June 19). *YouTube removes white supremacist content.* In *Weekend Edition Sunday.* NPR. https://www.npr.org/2019/06/09/731044416/youtube-removes-white-supremacist-content?t=1643586899974

Google. (n.d.-a). *Community Guidelines strike basics.* Retrieved December 26, 2021, from https://support.google.com/youtube/answer/2802032.

Google. (n.d.-b). *COVID-19 medical misinformation policy.* Retrieved December 26, 2021, from https://support.google.com/youtube/answer/9891785?hl=en.

Google. (n.d.-c). *Hate speech policy.* Retrieved December 27, 2021, from https://support.google.com/youtube/answer/2801939?hl=en.

Google. (n.d.-d). *Misrepresentation.* Retrieved December 27, 2021, from https://support.google.com/adspolicy/answer/6020955?hl=en.

Google. (n.d.-e). *Vaccine misinformation policy.* Retrieved December 26, 2021, from https://support.google.com/youtube/answer/11161123.

Google. (n.d.). *What happens if you violate our policies.* Advertising Policies Help. Retrieved December 27, 2021, from https://support.google.com/adspolicy/answer/7187501?hl=en&ref_topic=1308266.

Instagram. (n.d.). *COVID-19 and vaccine policy updates and protections.* Retrieved December 27, 2021, from https://help.instagram.com/697825587576762.

Meta. (n.d.-a). *Advertising policies related to coronavirus (COVID-19).* Retrieved January 4, 2022, from https://www.facebook.com/business/help/1123969894625935.

Meta. (n.d.-b). *Coordinating harm and promoting crime.* Retrieved December 26, 2021, from https://transparency.fb.com/en-gb/policies/community-standards/coordinating-harm-publicizing-crime/.

Meta. (n.d.-c). *Counting strikes.* Retrieved December 26, 2021, from https://transparency.fb.com/en-gb/enforcement/taking-action/counting-strikes/.

Meta. (n.d.-d). *Hate speech.* Retrieved December 27, 2021, from https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/.

Pinterest. (n.d.-a). *Account suspension.* Retrieved April 24, 2022, from https://help.pinterest.com/en/article/account-suspension.

Pinterest. (n.d.-b). *Community guidelines.* Retrieved April 26, 2022, from https://policy.pinterest.com/en/community-guidelines.

Rosen, G. (2020, April 16). An update on our work to keep people informed and limit misinformation about COVID-19. *Meta.* Retrieved January 4, 2022, from https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims.

Spotify. (2022, January 30). *Spotify platform rules.* https://newsroom.spotify.com/2022-01-30/spotify-platform-rules/

The YouTube Team. (2021, September 29). *Managing harmful vaccine content on YouTube.* YouTube Official Blog. https://blog.youtube/news-and-events/managing-harmful-vaccine-content-youtube/

TikTok. (2020a). *Community guidelines.* Retrieved December 27, 2021, from https://www.tiktok.com/community-guidelines?lang=en.

TikTok. (2020b). *Election integrity*. Retrieved December 26, 2021, from https://www.tiktok.com/safety/en-us/election-integrity/.

TikTok. (2021, February 24). *Community guidelines enforcement report*. https://www.tiktok.com/safety/resources/transparency-report-2020-2?lang=en

TikTok. (n.d.-a). *Content violations and bans*. Retrieved December 26, 2021, from https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans.

TikTok. (n.d.-b). *COVID-19*. Retrieved December 26, 2021, from https://www.tiktok.com/safety/en-us/covid-19/.

Twitter. (2020). *Coronavirus: Staying safe and informed on twitter*. Retrieved December 27, 2021, from https://blog.twitter.com/en_us/topics/company/2020/covid-19.

Twitter. (2021, January 8). *Permanent suspension of @realDonaldTrump*. https://blog.twitter.com/en_us/topics/company/2020/suspension

Twitter. (n.d.). *Abusive behavior*. Retrieved January 4, 2022, from https://help.twitter.com/en/rules-and-policies/abusive-behavior.

Twitter. (n.d.-a). *Civic integrity policy*. Retrieved December 26, 2021, from https://help.twitter.com/en/rules-and-policies/election-integrity-policy.

Twitter. (n.d.-b). *COVID-19 misleading information policy*. Retrieved December 26, 2021, from https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

YouTube. (n.d.). *How does YouTube support civic engagement and stay secure, impartial, and fair during elections?* Retrieved February 11, 2022, from https://www.youtube.com/intl/en_us/howyoutubeworks/our-commitments/supporting-political-integrity/.

YouTube Insider [@YoutubeInsider]. (2021, January 13). *After review, and in light of concerns about the ongoing potential for violence, we removed new content uploaded to Donald* [Tweet]. Twitter. https://twitter.com/YouTubeInsider/status/1349205688694812672?s=20&t=pMA3f60oCs6NI5ALZuI-Zw