SOCIAL SCIENCES

Partisan conflict over content moderation is more than disagreement about facts

Ruth E. Appel[1], Jennifer Pan[1], Margaret E. Roberts[2]*

Social media companies have come under increasing pressure to remove misinformation from their platforms, but partisan disagreements over what should be removed have stymied efforts to deal with misinformation in the United States. Current explanations for these disagreements center on the "fact gap"—differences in perceptions about what is misinformation. We argue that partisan differences could also be due to "party promotion"—a desire to leave misinformation online that promotes one's own party—or a "preference gap"—differences in internalized preferences about whether misinformation should be removed. Through an experiment where respondents are shown false headlines aligned with their own or the opposing party, we find some evidence of party promotion among Democrats and strong evidence of a preference gap between Democrats and Republicans. Even when Republicans agree that content is false, they are half as likely as Democrats to say that the content should be removed and more than twice as likely to consider removal as censorship.

INTRODUCTION

Misinformation is seen as a major global threat by political and economic leaders around the world (*1*) as well as by the general public (*2*, *3*). Rising public awareness of online misinformation has coincided with growing public debates about what social media companies should remove from their platforms. These debates have laid bare deep partisan divisions over the removal of online content in the United States. Both Republicans and Democrats have called for the repeal of Section 230 of the Communications Decency Act, which protects social media companies from liability for content on their platforms. But the two sides of the aisle have very different views about how the act should be reformed (*4*). This divide has led to partisan gridlock over policies to combat misinformation. For example, the Biden administration's creation of the Disinformation Governance Board under the Department of Homeland Security was paused by Republican objections over its mission just 3 weeks after its announcement (*5*). Partisan consensus over content moderation would empower social media companies to more effectively regulate what content should be permitted online. In contrast, conflict over content moderation puts both social media companies and regulators in a bind, as any decision is unpopular. Given that many large global social media platforms—such as Facebook, YouTube, WhatsApp, and Instagram—are based in the United States, U.S. content moderation policies may also influence content removal and moderation in other countries, some of which face their own partisan divisions (*6*). Similar to how EU regulation has a global impact in policy areas such as privacy (*7*), U.S. content moderation policies may also have global implications.

There is a large literature on content moderation (*8*, *9*), which has documented partisan differences in support for content removal (*10*, *11*). The most prominent explanation for this partisan disagreement over content moderation is what we call the "fact gap"—the idea that partisanship influences what individuals believe to be true (*12*, *13*) and hinders their ability to identify content aligned with their political views and ideology as misinformation (*14*–*21*). This fact gap could be driven by psychological mechanisms, including mechanisms for preserving one's identity (*22*, *23*) or complementary beliefs (*24*–*26*), such as motivated reasoning, prior attitude effect, and confirmation bias, as well as more general cognitive mechanisms like inattention (*27*–*29*).

Here, we theorize and test for the existence of two additional potential sources of partisan disagreement over what content social media companies should remove from the internet beyond the fact gap: "party promotion" and a "preference gap." Although the fact gap is important and consequential, it may not fully explain the sizeable partisan gaps that have been identified. Existing research suggests that analyzing factors beyond disagreement about facts may be important to understanding partisan disagreement over content moderation. For example, Kozyreva *et al.* (*10*) find a 30% partisan difference in preferences to remove content that denies the Holocaust and note that accuracy perceptions alone cannot explain such partisan differences.

We define party promotion as the desire to leave misinformation online when it benefits one's own party or denigrates the other party, and to remove misinformation that denigrates one's own party or promotes the other party, regardless of belief in the accuracy of the information. Partisanship might lead to such behavior due to the importance of the symbolic social standing of one's own party (in-group) relative to the other party (out-group) (*30*). In the United States, party promotion may be a plausible potential explanation for conflict over what content should be moderated given that affective polarization—the gap in affect toward the partisan in-group and the partisan out-group—has increased (*31*, *32*). Studies have identified phenomena similar to party promotion in the United States in related settings. For example, partisan alignment affects the demand for biased news (*33*) and predicts misinformation sharing (*9*). Content flagging has also been shown to be used strategically at times to promote one's own political aims rather than due to genuine belief (*34*).

Even in the absence of a fact gap or party promotion, partisans may disagree about whether content should be removed because of differences in internalized preferences. This preference gap implies

[1]Department of Communication, Stanford University, Stanford, CA 94305, USA.
[2]Department of Political Science and Halıcıoğlu Data Science Institute, University of California, San Diego, CA 92093, USA.
*Corresponding author. Email: meroberts@ucsd.edu

that there might be partisan differences in overall preferences for content removal on the internet, regardless of which party the specific content advantages and even if partisans agree that specific content is misinformation. This gap in preferences could stem from (a) differences in internal factors like identity or core values, which are deeply rooted and difficult to change, or (b) internalization of elite cues and signals, which may be more changeable. In terms of (a), people differ in identity, values, personality, cognitive processes, motives, and emotions (*22*, *23*, *35*–*41*). If people select into political parties based on these internal factors, then these divergent underlying factors could account for a partisan preference gap.

The preference gap could also result from (b), people internalizing cues and signals from elites in the party they identify with as their own preferences (*42*–*44*). Democratic and Republican elites have emphasized free speech as a core value in different periods of American history and on different issues (*11*, *45*, *46*) [see also section S2.3 for frequency of congressional speeches containing censorship-related keywords by party from the 46th (1879) to 116th U.S. Congress (2021)]. For example, Lynn Woolsey, a Democratic House member at the time, commented that increasing the concentration of media ownership could result in censorship in a 2003 speech (*47*). In early 2023, Republican House member Nicholas Langworthy expressed concern that Big Tech companies were censoring conservative voices (*48*). In recent years, Republican elites have framed online content removal as a free speech and censorship issue (*49*, *50*), while Democratic elites have generally been supportive of the need for content moderation. Note, however, that during this same time period, Democrats have expressed concern about censorship in other areas such as textbook bans [see speech by Democratic House member Jeremy Raskin in March 2023 (*51*)]. Such elite signaling may result in a preference gap because Republicans, knowing that party elites are opposed to content removal on the internet, may base their preferences on elite signals and prefer that content remains online, while Democrats, knowing that party elites support content moderation, may prefer removal of misinformation. Recent surveys show that Republicans place higher importance on free speech rights on the internet than Democrats, while Democrats place higher importance on preventing the spread of false information online than Republicans (*46*). Note that while this behavior appears similar to promoting one's own party, it reflects an internalized overall preference toward content moderation regardless of the partisan slant of the information. Thus, it differs from our concept of party promotion, which is strategic behavior that treats content online differently depending on its partisan slant.

Design and data

To test whether partisan conflict over content moderation may arise from the preference gap and party promotion, we embedded an experiment in a national survey of U.S. respondents. We attempt to neutralize the fact gap by presenting participants with misinformation headlines and explicitly telling respondents they are false. We then disaggregate the effects of the preference gap and party promotion by varying the partisan alignment of the headline.

Our survey of U.S. adults was commissioned by the Knight Foundation and fielded by Ipsos in the summer of 2021. The survey was implemented on the Ipsos KnowledgePanel, which is described by Ipsos as a representative random sample (for descriptive statistics comparing the sample to the U.S. population, see section S1.7). For our analysis, we focus on English-speaking respondents who identified as Democrat or Republican, resulting in 1120 respondents, with a mean age of 53.29 (SD = 16.53) and 56.3% female (see section S1.7 for detailed descriptive statistics). The experiment and analyses were preregistered (see data accessibility statement in the Acknowledgments for details; see deviations and clarifications from the pre-analysis plan in section S1.2 and throughout the supplementary materials text where they pertain). This research was approved by the Institutional Review Boards at our respective universities.

The survey experiment relied on simple randomization at the participant and at the headline level. Each participant was shown two different false news headlines sequentially (for a flow diagram of the experiment, see section S1.1). Respondents were told that "Someone has shared the following headline on a social media site. (This headline has been established as **false** by third-party fact checkers.)." One of the headlines aligned with the respondent's partisanship, while the other headline was not aligned with the respondent's partisanship. For example, one pro-Republican headline (aligned for Republicans, misaligned for Democrats) reads: "Hours after signing an executive order on Jan. 20, 2021, U.S. President Joe Biden violated his own mask mandate." Whether the respondent saw the aligned or misaligned headline first was randomized. Headlines were selected from a bank of 18 news headlines (9 aligned for Democrats, 9 aligned for Republicans) that contained false claims. We provide more information on headline selection in Materials and Methods.

We measured three main outcomes: (i) *Intent to remove headline (removal)*: Whether or not the participant states that the headline should be removed by the social media company; (ii) *Perception of headline removal as censorship (censorship)*: Whether the participant considers the removal of the headline censorship; (iii) *Intent to report headline as harmful (harm)*: Whether the participant would report the headline as harmful content on a social media platform. We also measure a range of covariates, including perceived accuracy. To measure perceptions of accuracy, we ask respondents for their perceived accuracy of the false news headlines on a four-point scale. All measures, including control variables and indices, are described in detail in section S1.6.

We analyze results using OLS regression, interacting partisanship of participants and political alignment of the headlines:

$$Y_{ia} = \beta_D D_i \cdot Hd_a + \beta_R R_i \cdot Hr_a + \gamma_D D_i + \gamma_R R_i + \varepsilon_{ia} \qquad (1)$$

where Y_{ia} is the binary outcome measure for individual i and headline a. D_i indicates that respondent i is a Democrat and R_i indicates that respondent i is a Republican. The difference in coefficients on D_i and R_i reflects the preference gap or the amount overall that Democrats and Republicans disagree about whether false content should be removed controlling for alignment. Hd_a is an indicator of whether headline a is aligned for Democrats and Hr_a is an indicator of whether headline a is aligned for Republicans. The coefficients on $D_i \cdot Hd_a$ and $R_i \cdot Hr_a$ reflect party promotion or the amount that the outcome depends on the alignment between the partisan nature of the content and the respondent for Democrats and Republicans, respectively (see section S1.3 for additional details on our analyses).

RESULTS

We find a large and statistically significant difference between the content moderation preferences of Republicans and Democrats. Overall, the probability that Democrats say a false headline should be removed is 0.69, while the probability that Republicans say a false headline should be removed is 0.34. The probability that Democrats would report a false headline as harmful is 0.49, while for Republicans, it is 0.27. The probability that Democrats perceive the removal of false headlines as censorship is 0.29, while for Republicans, it is 0.65 (see tables S12 to S14 for regressions that calculate these probabilities).

The left panels of Fig. 1 plot the coefficient estimates and confidence intervals from Eq. 1 for all respondents and each of the three outcomes. The right panels of Fig. 1 present the same estimates along with the overall gap between partisans to illustrate the relative sizes of party promotion and the preference gap (see section S1.3.1 for details). The preference gap for the removal outcome is the difference between Democrats' and Republicans' support for removal, controlling for alignment. Party promotion for the removal outcome is the difference between Democrats' and Republicans' support for the removal of aligned versus misaligned headlines.

Looking at the intent to remove headline outcome (Fig. 1, A and B), we can see that for misaligned headlines, the probability of intent to remove is 0.75 for Democrats, while it is 0.34 for Republicans, resulting in a misaligned preference gap of 0.41. For Republicans, there is no difference in their intent to remove misaligned and aligned headlines (no party promotion), but there is party promotion for Democrats (intent to remove aligned headlines declines by 0.11 to 0.64). For the intent to report headline as harmful outcome (Fig. 1, C and D), we also see a sizable preference gap between Democrats and Republicans (0.30 preference gap for misaligned headlines), some party promotion among Democrats who are less likely (−0.13) to report aligned headlines as harmful than misaligned headlines, and no party promotion among Republicans who are equally willing to report aligned and misaligned headlines as harmful. For the perception of headline removal as a censorship outcome (Fig. 1, E and F), there is no evidence of party promotion —no difference between misaligned versus aligned headlines among Democrats or Republicans—but a large preference gap between Democrats and Republicans (−0.37 for misaligned headlines).

Persistence of the fact gap

While we inform respondents that the headlines have been rated as false by third-party fact-checkers, respondents rated 20.32% of headlines as either "very accurate" or "somewhat accurate." Moreover, consistent with previous literature, the interaction terms in Fig. 2 show that evaluations of the accuracy of the headline are partisan—both Democrats and Republicans are more likely to think that headlines that align with their own position are true, reflecting the persistence of the fact gap despite explicit information we provided that the headlines are false. From Fig. 2, we see that Democrats rate 11% of pro-Republican and 25% (11% + 14%) of pro-Democrat headlines as accurate. Similarly, Republicans rate 21% of pro-Democrat headlines and 32% (21% + 11%) of pro-Republican headlines as accurate.

If some respondents still believed that the headlines were true, despite being told they were false, this poses a problem of identification. It means that we cannot fully isolate the effects of party promotion and the preference gap from the effect of the fact gap. To address this, we conducted three additional analyses, two of which were preregistered. If the results of all three tests are consistent, then it would give us further confidence that we can measure the contribution of party promotion and the preference gap to partisan disagreement over content moderation, controlling for the fact gap. For more information on each method, see Materials and Methods.

Inaccurate subgroup analysis

In the first analysis to address this persistent fact gap, we subset to respondents who rated the headlines as inaccurate, including respondents who assessed headlines as "not very accurate" or "not at all accurate." Figure 3 shows that when we subset to respondents who agree the false headlines are inaccurate, the preference gap results stay the same. Democrats are still nearly twice as likely as Republicans to want to remove the headline and report the headline as harmful, and half as likely to perceive removal as censorship. While Republican respondents still exhibit the same preferences on all three outcomes regardless of whether the false headline is aligned or misaligned with their political views, party promotion among Democrats is slightly smaller among the inaccurate subgroup. This suggests that some of the party promotion in the main results may have been a result of the fact gap. However, there is still a significant effect of party promotion among Democrats, suggesting that factual beliefs do not completely explain away this effect.

A shortcoming of this approach, which means we cannot fully account for the fact gap, is that Democrats and Republicans who believe the headline to be inaccurate are potentially differentially selected. We find that on observables such as gender, age, income, and education, Democrats and Republicans who assess headlines as inaccurate are not significantly different from Democrats and Republicans in the overall sample (see tables S3 and S4). However, there could still be differences in unobservables between those who assess headlines to be inaccurate and the broader respondent pool.

Consensus headlines analysis

In the second analysis, we conducted the main analyses for headlines that, on average, both Republicans and Democrats think are inaccurate and where there is little difference in accuracy perception between Democrats and Republicans (see section S2.1.5). We identify these "consensus headlines" by limiting the mean absolute difference between the average Democrat and Republican accuracy rating for headlines that both Democrats and Republicans on average rate as inaccurate to 0.5 on the four-point accuracy scale. This results in eight headlines. When we decrease this threshold, reducing the number of headlines, the substantive results remain unchanged; see section S2.1.5. We add this analysis to attempt to address the concern that the gap in support for removal observed among Democrats and Republicans in the previous analyses is driven by headlines with a large gap in perceived accuracy between Democrats and Republicans (see also results disaggregated by headlines in section S2.1.6).

While a limitation of this robustness check is that it selects only very few headlines and therefore may not generalize, at least for this set of headlines, we continue to observe evidence for the preference gap when we hone in on headlines Democrats and Republicans

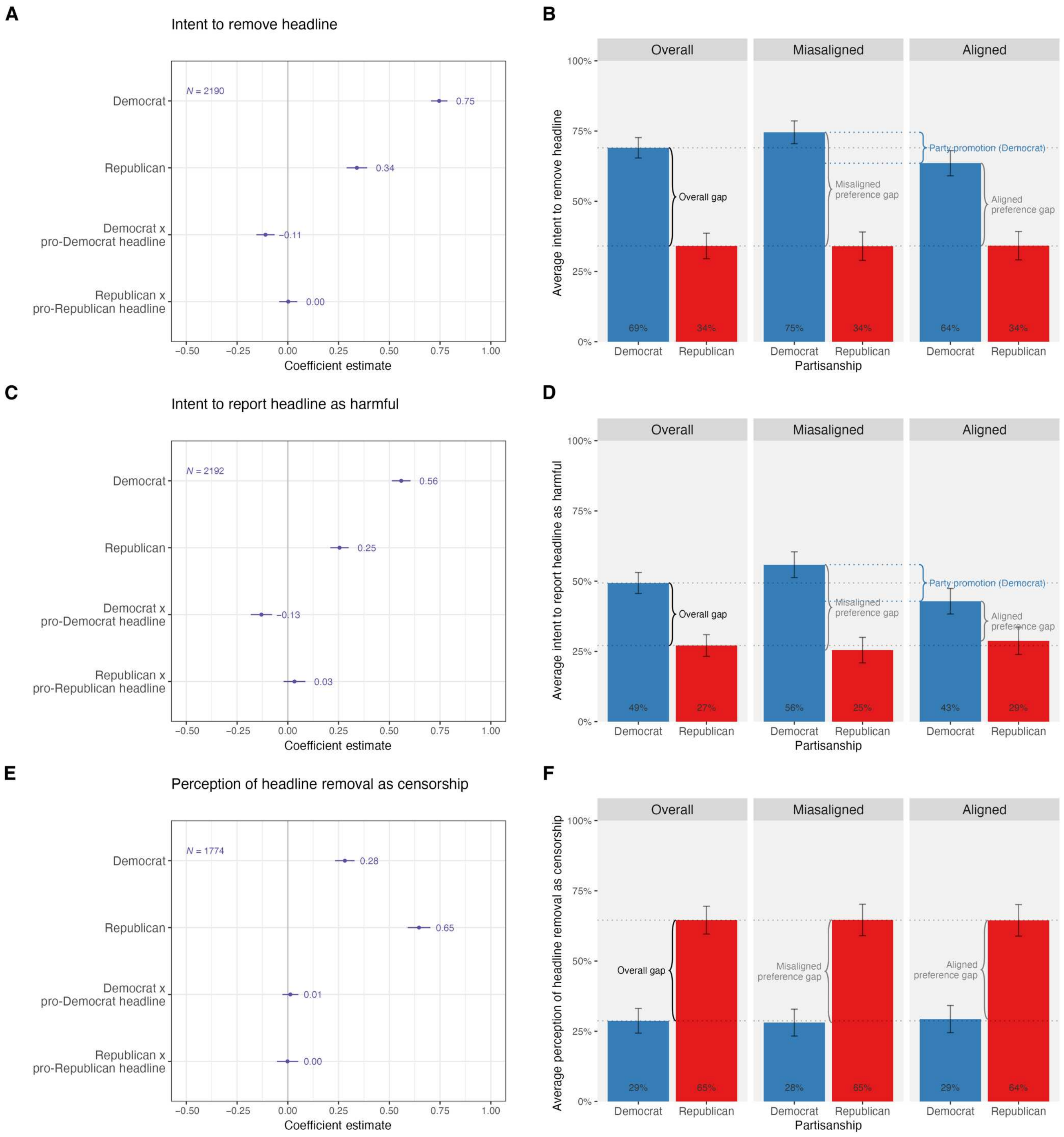

Fig. 1. Partisanship and preferences for content moderation for all respondents. No control variables; 95% confidence intervals are shown. The left panels show the coefficient estimates and confidence intervals from Eq. 1 for all respondents and the removal (**A**), harm (**C**), and censorship (**E**) outcome. The right panels show the same estimates for all respondents along with the overall gap between partisans for the removal (**B**), harm (**D**), and censorship (**F**) outcome (see section S1.3.1 for details).

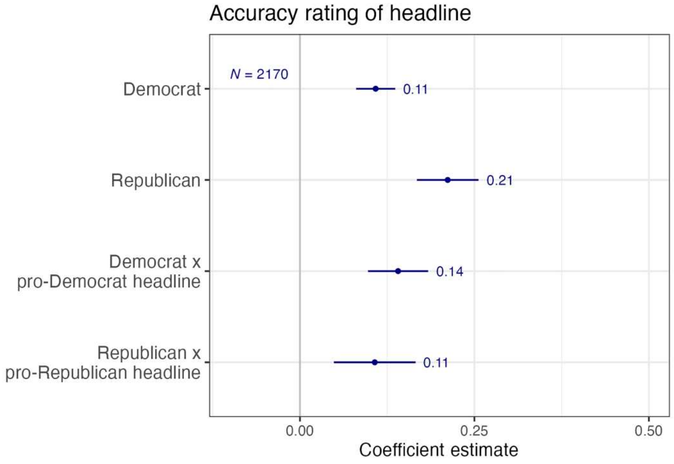

Fig. 2. Respondents' assessment of headline accuracy. No control variables.

agree are inaccurate. Democrats remain nearly twice as likely as Republicans to want to remove content and to report content as harmful, while Republicans are nearly twice as likely as Democrats to consider removal censorship. We also continue to see evidence of party promotion among Democrats.

Accuracy as mediator

In the third analysis, we examine the extent to which party promotion among Democrats and the preference gap are mediated by belief in the accuracy of the content. For the mediation effect for party promotion, we conducted a mediation analysis of the effect of alignment between respondent and headline partisanship on the outcomes for Democratic respondents. The mediation analysis was preregistered for party promotion. We also conducted a mediation analysis for the preference gap, which was not preregistered. For the mediation effect of the preference gap, we conducted a mediation analysis of the effect of Democrat partisanship on the outcomes for all respondents. For more details on the mediation analyses, see sections S2.2.1 and S2.2.2.

Table 1 shows the results of the analysis for party promotion. The estimand in this analysis is the average causal mediation effect (ACME) (*52*). ACME is the total effect that alignment has on the outcome variable of interest minus the average direct effect (ADE), which is the effect of alignment on the outcome without taking the indirect path through accuracy into account.

In the main analysis (Fig. 1), we saw that Democrats were less likely to intend to remove a headline or report a headline as harmful for headlines aligned with their partisanship. In Table 1, we see that this effect may, in part, be mediated by accuracy. ACME and ADE are negative and significant for both intent to remove the headline and intent to report the headline as harmful (for alternative specifications, see tables S25 and S26). This suggests that, while the party promotion effect for Democrats is reduced when accounting for perceptions of accuracy, the fact gap may not completely explain away party promotion among Democrats on these outcomes.

Note that ACME is only identified under a sequential ignorability assumption that (i) given the observed pretreatment confounders, the treatment assignment is statistically independent of potential outcomes and potential mediators and (ii) the mediator is ignorable given observed treatment and pretreatment confounders. While treatment is randomly assigned, the second assumption may not hold because the mediator, perception of accuracy, is not. In other words, there may be characteristics of respondents that affect both whether they think a headline is accurate and whether they think a headline should be removed. To probe the robustness of the findings, we conducted a sensitivity analysis (*53*). The sensitivity analysis indicates that our conclusion, that accuracy mediates party promotion but cannot completely explain away party promotion for Democrats, is plausible given even fairly large departures from the ignorability of the mediator due to pretreatment confounders (see section S2.2 for additional details).

As described in section S2.2.2, we also conduct a mediation analysis for the preference gap. We find that the preference gap is largely a direct effect of partisanship on content moderation preferences, only mediated by accuracy by a small amount. We also conducted a sensitivity analysis and find that this direct effect is robust to very large departures from the ignorability of the mediator due to pretreatment confounders.

We acknowledge that partisan disagreement on headline accuracy poses a problem of identification. However, the results of the three additional analyses that we conducted to address the fact gap are consistent and suggest that the preference gap and, to a smaller extent, party promotion explain a portion of partisan differences in content moderation preferences.

DISCUSSION

In line with existing research on content moderation, we find strong partisan differences in content moderation preferences. However, the results of this experiment highlight a need to consider factors beyond the fact gap to understand these partisan differences. While prior research has established the importance of the fact gap in explaining content moderation preferences, our experiment shows that the preference gap likely also affects attitudes toward the removal of misinformation online. In the United States today, Democrats prefer to remove misinformation, while Republicans prefer to avoid the removal of misinformation and perceive such removal as censorship, even when they agree that the content is inaccurate. In addition, this study provides a previously unexplored perspective on debates related to content moderation because although the term "censorship" is used in major political debates in the United States, most studies related to censorship perceptions were conducted before the social media era [for examples and exceptions, see (*22*, *45*, *54*, *55*)].

There are limitations to this study that highlight the need for future research to examine the causes of the preference gap, to study this dynamic at other points in time and in other political contexts, to measure how these findings generalize from a survey experiment to social media platforms, and to look at content beyond the political misinformation examined here. The preference gap could arise from deeply rooted internal factors such as moral values or from internalization of elite cues. Differentiating between these factors has important policy implications because internal factors like moral values are difficult to move, while elite cues may be more likely to change over time. Fifteen years ago, Lindner and Nosek (*11*) found that Democrats had stronger preferences for protecting free speech than Republicans, perhaps suggesting that more changeable factors may be at work, but additional empirical

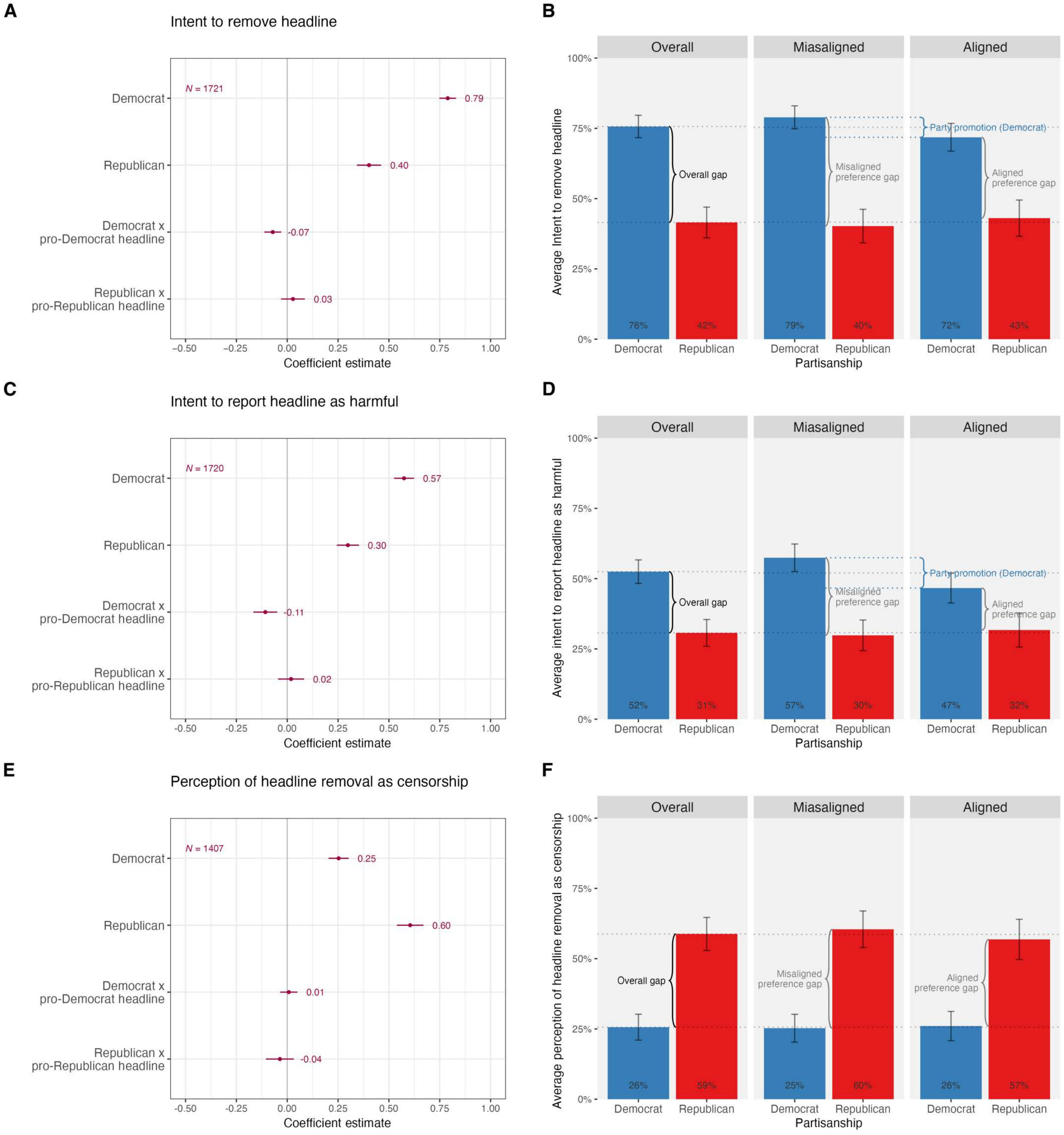

Fig. 3. Partisanship and preferences for content moderation for respondents who agree headlines are inaccurate. No control variables; 95% confidence intervals are shown. The left panels show the coefficient estimates and confidence intervals from Eq. 1 for respondents who agree headlines are inaccurate and the removal (**A**), harm (**C**), and censorship (**E**) outcome. The right panels show the same estimates for respondents who agree headlines are inaccurate along with the overall gap between partisans for the removal (**B**), harm (**D**), and censorship (**F**) outcome (see section S1.3.1 for details).

Table 1. Effect of alignment mediated by accuracy for Democrats. Note: Mediation models were run with standard errors clustered on participants, without weighting observations and without control variables using a dataset in which missing values were addressed using listwise deletion. ACME, average causal mediation effect; ADE, average direct effect.

Measure	Estimate	*P* Value
Intent to remove headline		
ACME	−0.065	<0.001
ADE	−0.039	0.034
Total effect	−0.103	<0.001
Proportion mediated	0.624	<0.001
N Observations	1302	
N Simulations	1000	
Intent to report headline as harmful		
ACME	−0.035	<0.001
ADE	−0.074	<0.001
Total effect	−0.109	<0.001
Proportion mediated	0.321	<0.001
N Observations	1301	
N Simulations	1000	

assessments are needed. This underscores the need to study these dynamics at different points in time and in other political contexts. For example, future research could pursue a similar experiment in contexts where those on the right are more supportive of content moderation and those on the left oppose it (*56*).

In our experiment, we balanced the partisanship of headlines and kept other headline characteristics and their context relatively comparable. On social media platforms, however, Republican-aligned misinformation is more common (*57*). It could be that this difference in the prevalence of misinformation drives differences in the content moderation preferences of Democrats and Republicans. For example, Republicans might have developed lower baseline moderation preferences because they think that the content moderation system disproportionately targets them. Alternatively, if Republicans' threshold for unfollowing users mimics their high threshold for removing content, then our findings could also explain why conservatives are exposed to more misinformation in general (*58*–*62*). Future research could explore this in more detail.

Experimenter demand effects, social desirability bias, and the higher cost and benefit of taking action in a real-world setting could affect our results. While the action of flagging content on social media is similar to clicking on a button in a survey, users may experience such actions differently on social media knowing that their actions may have tangible, real-world consequences. In this study, we did not use incentivized responses to address these issues. This experiment was part of a larger collaborative survey, thus we did not have the opportunity to provide incentives. Furthermore, there is debate over how incentivized responses influence studies of partisan differences (*63*), with some finding reductions in partisan differences (*13*, *64*) and others arguing that such designs are not required to capture how people in the real world evaluate and make decisions (*16*, *65*). Future work could explore to which extent our results generalize beyond a survey context, for example, by embedding a similar experiment within a social media platform or adding treatment groups with incentivized responses such as a willingness-to-pay design.

Last, the content (e.g., political versus health misinformation) and context (e.g., motivation to seek out the truth, how rooted beliefs about a topic are in one's identity) of misinformation headlines also matter (*10*, *22*, *23*, *46*, *66*). In this study, we focus on political headlines, and more specifically those denigrating out-party politicians instead of flattering in-party politicians. Future research should investigate further how other types of content—political misinformation flattering in-party politicians, nonpolitical misinformation, hate speech, voter suppression content—and different contexts influence the preference gap and party promotion (*67*). Another potentially interesting research question is to what extent individual-level drivers of content moderation are decisive at the level of content moderation systems with thousands of often professional content moderators, and which other factors might be at play in those systems.

In terms of the implications of these findings, it is encouraging that the effects of party promotion are dwarfed by the preference gap. In an environment with increasing partisan animosity, respondents—Republicans in particular—seemed to evaluate content removal outside of the lens of party promotion. Policymakers and social media platforms could consider different approaches to design policies with bipartisan support. First, thinking about content moderation as a system of procedures applied at scale, rather than decisions on individual pieces of content by individual moderators (*68*), might help by shifting the focus from specific content to be moderated to a system of procedures that needs to be agreed upon. For this system, the preference gap might be less pronounced than for specific content. Second, future research could explore whether there might be a partisan consensus on less extreme forms of content moderation, like flagging or downweighting misinformation. Third, policymakers could attempt to use moral reframing, the practice of tailoring content to an individual's moral values by framing a position an individual would usually oppose in a way that is consistent with their moral values (*69*), to bridge the preference gap to the extent that it is rooted in moral value differences.

Policymakers and social media platforms should understand that differences between Democrats and Republicans stem from more than just disagreement over what is true versus false and strategic partisan maneuvering. Instead, Americans seem to have diverging preferences about the concept of content removal and whether the protection of free speech necessitates or precludes the moderation of content.

MATERIALS AND METHODS

Sample

Our sample consisted of U.S. adults recruited by Ipsos, which is a market research firm based in France with worldwide operations. Ipsos is nonpartisan and we have no indication that U.S. respondents perceived it as having biases, ideological or otherwise. Following the exclusion criteria laid out in our pre-analysis plan, we only included participants who indicated that their preferred language was English and excluded participants who self-identified as independents because the alignment treatment would not work for independents. As noted in section S1.2, the pre-analysis plan did not specify how we would address unexpected missing values or participants from a different sample, thus we added clarifications for the following actions: We excluded participants who had missing values for partisanship or indicated values other than Democrat, Republican, or independent. Another 243 participants were part of a student sample that was different from the sample meant to be representative of the U.S. population, and we therefore excluded them from our analysis. The data were weighted with the weights provided by Ipsos for the models presented in the main text, but we also report unweighted results in section S2. See section S1.7 for detailed descriptive statistics.

Headline selection

We identified headlines used in the experiment from the fact-checking website Snopes.com with the criteria that the headline included a clear "false" label (not partially or entirely true), political content, a clear partisan slant, and was recent. We selected headlines that were relatively balanced in terms of the intensity of the information conveyed (e.g., level of violence) and the topic. We then used a pretest to ensure that candidate headlines had a partisan alignment in the expected direction and to measure other headline characteristics including the perceived intensity (e.g., how worrying the headline is). Although these headlines may differ in other dimensions that we did not assess and may not be perfectly comparable, the final selection of headlines included pro-Democrat and pro-Republican headlines with the expected ideological slant that were relatively balanced in terms of perceived intensity and topic (see section S1.5 for additional details).

Outcome measure details

All outcome measures are binary with the exception of the censorship measure, which was recoded as a binary measure by considering "Yes" as 1, "No" as 0, and "Do not know" as a missing value. We deviated from the pre-analysis plan in recoding "Do not know" as a missing value instead of 0 because recoding "Do not know" as 0 would have imposed a strong assumption that undecided participants actually did not think of headline removal as censorship. We provide results for the main models with the original coding as a robustness check in the Supplementary Materials (see section S2.1.2), and find that the main results remain the same.

Accuracy measure

To measure perceptions of accuracy, we asked respondents for their perceived accuracy of the false news headlines on a four-point scale. We randomized whether participants first answered the question about accuracy or the outcome questions after the headline throughout the experiment (see fig. S1). This also allows us to ensure that our results were not driven by an "accuracy nudge" (*28*, *29*). Participants had a 50% chance of being asked the perceived accuracy question before any outcome variables were measured and a 50% chance of being asked the perceived accuracy question after the outcome variables of removal and censorship were measured (see section S1.4.2 for balance tables, section S2.1.3 for analyses with the first headline only, and tables S22 to S24 for the accuracy order analysis). The measure for harm was always asked last because we did not want to influence accuracy ratings by priming participants to think about harm.

Covariates and indices

We measured a range of covariates such as news consumption and demographics as detailed in section S1.6. Control variables include age, gender, education, race, ethnicity, household income, political interest, whether social media was the most common news source, and whether a participant's posts had ever been flagged or removed from social media. The order of response options in several questions on covariates, such as partisanship, was randomized. Some of the control variables that we include in our regressions are measured by multiple survey questions. For such questions, we used composite indices as detailed in section S1.6.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S21
Tables S1 to S28
References

REFERENCES AND NOTES

1. World Economic Forum, "The global risks report 2023" (Tech. Rep. 18, World Economic Forum, 2023).
2. Pew Research Center, "Climate change remains top global threat across 19-country survey" (Tech. Rep., Pew Research Center, 2022).
3. L. Silver, "Americans see different global threats facing the country now than in March 2020" (Tech. Rep., Pew Research Center, 2022).
4. D. E. Bambauer, *What Does the Day After Section 230 Reform Look Like?* (Brookings Institution, 2021).
5. T. Lorenz, *How the Biden Administration Let Right-Wing Attacks Derail Its Disinformation Efforts* (The Washington Post, 2022).
6. K. Ruggeri, B. Većkalov, L. Bojanić, T. L. Andersen, S. Ashcroft-Jones, N. Ayacaxli, P. Barea-Arroyo, M. L. Berge, L. D. Bjørndal, A. Bursalıoğlu, V. Bühler, M. Čadek, M. Çetinçelik, G. Clay, A. Cortijos-Bernabeu, K. Damnjanović, T. M. Dugue, M. Esberg, C. Esteban-Serna, E. N. Felder, M. Friedemann, D. I. Frontera-Villanueva, P. Gale, E. Garcia-Garzon, S. J. Geiger, L. George, A. Girardello, A. Gracheva, A. Gracheva, M. Guillory, M. Hecht, K. Herte, B. Hubená, W. Ingalls, L. Jakob, M. Janssens, H. Jarke, O. Kácha, K. N. Kalinova, R. Karakasheva, P. R. Khorrami, Ž. Lep, S. Lins, I. S. Lofthus, S. Mamede, S. Mareva, M. F. Mascarenhas, L. M. Gill, S. Morales-Izquierdo, B. Moltrecht, T. S. Mueller, M. Musetti, J. Nelsson, T. Otto, A. F. Paul, I. Pavlović, M. B. Petrović, D. Popović, G. M. Prinz, J. Razum, I. Sakelariev, V. Samuels, I. Sanguino, N. Say, J. Schuck, I. Soysal, A. L. Todsen, M. R. Tünte, M. Vdovic, J. Vintr, M. Vovko, M. A. Vranka, L. Wagner, L. Wilkins, M. Willems, E. Wisdom, A. Yosifova, S. Zeng, M. A. Ahmed, T. Dwarkanath, M. Cikara, J. Lees, T. Folke, The general fault in our fault lines. *Nat. Hum. Behav.* **5**, 1369–1380 (2021).
7. A. Bradford, The brussels effect. *Northwest. Univ. Law Rev.* **107**, 1–68 (2012).
8. R. Jiménez Durán, "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter." George J. Stigler Center for the Study of the Economy & the State Working Paper No. 324, University of Chicago, Chicago, 7 November 2023.
9. E. Henry, E. Zhuravskaya, S. Guriev, Checking and Sharing Alt-Facts. *Am. Econ. J. Econ. Policy.* **14**, 55–86 (2022).
10. A. Kozyreva, S. M. Herzog, S. Lewandowsky, R. Hertwig, P. Lorenz-Spreen, M. Leiser, J. Reifler, Resolving content moderation dilemmas between free speech and harmful misinformation. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2210666120 (2023).
11. N. M. Lindner, B. A. Nosek, Alienable speech: Ideological variations in the application of free-speech principles. *Polit. Psychol.* **30**, 67–92 (2009).
12. J. G. Bullock, G. Lenz, Partisan bias in surveys. *Annu. Rev. Polit. Sci.* **22**, 325–342 (2019).
13. M. Prior, G. Sood, K. Khanna, You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quart. J. Polit. Sci.* **10**, 489–518 (2015).
14. H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236 (2017).
15. C. Batailler, S. M. Brannon, P. E. Teas, B. Gawronski, A signal detection approach to understanding the identification of fake news. *Perspect. Psychol. Sci.* **17**, 78–98 (2022).
16. D. Flynn, B. Nyhan, J. Reifler, The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Polit. Psychol.* **38**, 127–150 (2017).

17. M. Jakesch, M. Koren, A. Evtushenko, M. Naaman, The role of source, headline and expressive responding in political news evaluation. https://ssrn.com/abstract=3306403 (2018).

18. C. S. Traberg, S. van der Linden, Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personal. Individ. Differ.* **185**, 111269 (2022).

19. S. C. Rhodes, Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation. *Polit. Commun.* **39**, 1–22 (2022).

20. J. Roozenbeek, R. Maertens, S. M. Herzog, M. Geers, R. Kurvers, S. Mubashir, S. Van Der Linden, Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgm. Decis. Mak.* **17**, 547–573 (2022).

21. B. Gawronski, Partisan bias in the identification of fake news. *Trends Cogn. Sci.* **25**, 723–724 (2021).

22. A. Ashokkumar, S. Talaifar, W. T. Fraser, R. Landabur, M. Buhrmester, Á. Gómez, B. Paredes, W. B. Swann Jr., Censoring political opposition online: Who does it and why. *J. Exp. Soc. Psychol.* **91**, 104031 (2020).

23. J. J. Van Bavel, A. Pereira, The partisan brain: An identity-based model of political belief. *Trends Cogn. Sci.* **22**, 213–224 (2018).

24. Z. Kunda, The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).

25. C. G. Lord, L. Ross, M. R. Lepper, Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* **37**, 2098–2109 (1979).

26. C. S. Taber, M. Lodge, Motivated skepticism in the evaluation of political beliefs. *Am. J. Polit. Sci.* **50**, 755–769 (2006).

27. U. K. H. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, M. A. Amazeen, The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* **1**, 13–29 (2022).

28. G. Pennycook, D. G. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).

29. G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, D. G. Rand, Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).

30. L. Huddy, *The Oxford Handbook of Political Psychology*, L. Huddy, D. Sears, J. Levy, Eds. (Oxford Univ. Press, ed. 2, 2013), pp. 737–773.

31. S. Iyengar, S. J. Westwood, Fear and loathing across party lines: New evidence on group polarization. *Am. J. Polit. Sci.* **59**, 690–707 (2015).

32. S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, S. J. Westwood, The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).

33. F. Chopra, I. Haaland, C. Roth, "The demand for news: Accuracy concerns versus belief confirmation motives," Working Paper 01/2023, NHH Department of Economics, 12 January 2023.

34. K. Crawford, T. Gillespie, What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media Soc.* **18**, 410–428 (2016).

35. M. Feinberg, R. Willer, From gulf to bridge. *Pers. Soc. Psychol. Bull.* **41**, 1665–1681 (2015).

36. L. Silver, P. van Kessel, "Both Republicans and Democrats prioritize family, but they differ over other sources of meaning in life" (Tech. Rep., Pew Research Center, 2021).

37. A. Campbell, P. E. Converse, W. E. Miller, D. E. Stokes, *The American Voter* (Univ. of Chicago Press, 1980).

38. J. Duckitt, C. G. Sibley, Personality, ideology, prejudice, and politics: A dual-process motivational model. *J. Pers.* **78**, 1861–1894 (2010).

39. P. Goren, Party identification and core political values. *Am. J. Polit. Sci.* **49**, 881–896 (2005).

40. J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**, 1029–1046 (2009).

41. J. T. Jost, C. M. Federico, J. L. Napier, Political ideology: Its structure, functions, and elective affinities. *Annu. Rev. Psychol.* **60**, 307–337 (2009).

42. C. Ellis, J. A. Stimson, *Ideology in America* (Cambridge Univ. Press, 2012).

43. M. Fiorina, S. Abrams, J. Pope, *Culture War? The Myth of a Polarized America* (Pearson Longman, 2005).

44. N. McCarty, K. T. Poole, H. Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches* (MIT Press, 2016).

45. R. D. Fisher, S. Lilie, C. Evans, G. Hollon, M. Sands, D. DePaul, C. Brady, D. Lindbom, D. Judd, M. Miller, T. Hultgren, Political ideologies and support for censorship: Is it a question of whose ox is being gored? *J. Appl. Soc. Psychol.* **29**, 1705–1731 (1999).

46. Knight Foundation, Ipsos, "Free expression in America post-2020" (Tech. Rep., Knight Foundation, 2022).

47. The growing concentration of media ownership. *Congressional Record* **149**, H4179–H4184 (2003).

48. Censorship of conservative voices from big tech corporations. *Congressional Record* **169**, H614–H615 (2023).

49. Rubio Introduces Sec 230 Legislation to Crack Down on Big Tech Algorithms and Protect Free Speech (2021).

50. Governor Ron DeSantis Signs Bill to Stop the Censorship of Floridians by Big Tech (2021).

51. Parents bill of rights act. *Congressional Record* **169**, H1348–H1383 (2023).

52. K. Imai, L. Keele, D. Tingley, A general approach to causal mediation analysis. *Psychol. Methods* **15**, 309–334 (2010).

53. K. Imai, L. Keele, T. Yamamoto, Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25**, 51–71 (2010).

54. R. Hense, C. Wright, The development of the attitudes toward censorship questionnaire. *J. Appl. Soc. Psychol.* **22**, 1666–1675 (1992).

55. J. T. Crawford, J. M. Pilanski, Political Intolerance, Right and Left. *Polit. Psychol.* **35**, 841–851 (2014).

56. J. Esberg, Censorship as reward: Evidence from pop culture censorship in Chile. *Am. Polit. Sci. Rev.* **114**, 821–836 (2020).

57. A. Rao, F. Morstatter, K. Lerman, Partisan asymmetries in exposure to misinformation. *Sci. Rep.* **12**, 15671 (2022).

58. G. Eady, T. Paskhalis, J. Zilinsky, R. Bonneau, J. Nagler, J. A. Tucker, Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nat. Commun.* **14**, 62 (2023).

59. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).

60. A. M. Guess, B. Nyhan, J. Reifler, Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).

61. M. Mosleh, D. G. Rand, Measuring exposure to misinformation from political elites on Twitter. *Nat. Commun.* **13**, 7144 (2022).

62. D. Nikolov, A. Flammini, F. Menczer, *Right and Left, Partisanship Predicts (Asymmetric) Vulnerability to Misinformation*. (Harvard Kennedy School Misinformation Review, 2021); https://doi.org/10.37016/mr-2020-55.

63. O. Yair, G. A. Huber, How robust is evidence of partisan perceptual bias in survey responses? *Public Opin. Q.* **84**, 469–492 (2021).

64. J. G. Bullock, A. S. Gerber, S. J. Hill, G. A. Huber, Partisan bias in factual beliefs about politics. *Quart. J. Polit. Sci.* **10**, 519–578 (2015).

65. A. J. Berinsky, Telling the truth about believing the lies? Evidence for the limited prevalence of expressive survey responding. *J. Polit.* **80**, 211–224 (2018).

66. G. Pennycook, D. G. Rand, Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022).

67. F. Pradel, J. Zilinsky, S. Kosmidis, Y. *Theocharis, Do Users Ever Draw a Line? Offensiveness and Content Moderation Preferences on Social Media* (OSF, 2022).

68. E. Douek, Content moderation as systems thinking. *Harv. Law Rev.* **136**, 526–607 (2022).

69. M. Feinberg, R. Willer, Moral reframing: A technique for effective and persuasive communication across political divides. *Soc. Personal. Psychol. Compass.* **13**, e12501 (2019).

70. J. Honaker, G. King, M. Blackwell, Amelia II: A program for missing data. *J. Stat. Softw.* **45**, 1–47 (2011).

71. M. Mosleh, C. Martel, D. Eckles, D. G. Rand, Perverse consequences of debunking in a Twitter field experiment: Being corrected for posting false news increases subsequent sharing of low quality, partisan, and toxic content, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 8 to 13 May 2021 (ACM, 2021), pp. 1–13.

72. D. Tingley, T. Yamamoto, K. Hirose, L. Keele, K. Imai, mediation: R Package for causal mediation analysis. *J. Stat. Softw.* **59**, 1–38 (2014).

73. D. Card, S. Chang, C. Becker, J. Mendelsohn, R. Voigt, L. Boustan, R. Abramitzky, D. Jurafsky, Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2120510119 (2022).

74. D. Card, S. Chang, C. Becker, J. Mendelsohn, R. Voigt, L. Boustan, R. Abramitzky, D. Jurafsky, Replication code and data for "Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration" [Dataset]; https://github.com/dallascard/us-immigration-speeches/ (2022).

75. History, Art & Archives, U.S. House of Representatives, *Party Government Since 1857* [Dataset] (History, Art & Archives, U.S. House of Representatives, 2022); https://history.house.gov/Institution/Presidents-Coinciding/Party-Government/.

Acknowledgments: We thank P. Barberá, A. Guess, E. Hersh, Y. Krupnikov, L. Opolka, J. Roozenbeek, Y. Theocharis, and participants of the NYU CSMaP Conference, the Stanford Polarization and Social Change Lab workshop, the Stanford Center on Philanthropy and Civil Society workshop, and the IC2S2 conference for helpful comments; A. Liria Sala for research assistance; and the Knight Foundation and Ipsos for supporting this work. The source of our data is Knight Foundation-Ipsos (2022) "Free Expression in America Post-2020" [Dataset]. **Funding:** R.E.A. has been supported by an SAP Stanford Graduate Fellowship in Science and Engineering, a Stanford Center on Philanthropy and Civil Society PhD Research Fellowship, a Stanford Impact Labs Summer Collaborative Research Fellowship, and the Stanford Center for American Democracy. **Author contributions:** R.E.A., M.E.R., and J.P. contributed to conceptualization, investigation, methodology, project administration, resources, supervision, validation, visualization, writing of the original draft, and review and editing. R.E.A. contributed to data curation, formal analysis, and software. **Competing interests:** R.E.A. has interned at Google in 2020; J.P. has received direct research funding from Meta in 2017 and received an honorarium from Meta for serving on the News Integrity Expert Circle from 2020-2021; R.E.A., J. P., and M.E.R. have attended an event where food was paid for by a social media company. The authors declare they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The experiment and analyses were preregistered on OSF at https://doi.org/10.17605/osf.io/3q4yn. Replication code and data are available on the Harvard-MIT Dataverse at https://doi.org/10.7910/dvn/um4vaj.

Submitted 18 January 2023
Accepted 3 October 2023
Published 3 November 2023
10.1126/sciadv.adg6799