Partisan conflict over content moderation is more than disagreement about facts*

Ruth E. Appel Jennifer Pan Margaret E. Roberts

20th January 2023

Abstract

Social media companies have come under increasing pressure to remove misinformation from their platforms, but disagreements between Republicans and Democrats over what should be removed have stymied efforts to deal with misinformation in the United States. In this paper, we identify three potential sources of partisan disagreement: 1) a "fact gap" – differences in perceptions about what is misinformation; 2) a "value gap" – differences in overall preferences about the amount of content that should be removed; and 3) "party promotion" – a desire to leave misinformation online that promotes one's own party. We conduct a survey experiment in a national survey of U.S. respondents that controls for the first factor and disaggregates the effects of the remaining two. We explicitly tell respondents that the content presented to them is misinformation and vary whether that content aligns with the respondent's party or the opposing party. We find strong evidence for a value gap. 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. While we find some evidence of Democrats' willingness to use content moderation for party promotion, overwhelmingly our results show that disagreement between Republicans and Democrats about content moderation comes from differences in values rather than strategic considerations of party promotion. These findings have important implications for policymakers and suggest that settling factual disagreements will not resolve partisan conflict over content moderation.

Keywords: social media, partisanship, censorship, content moderation

*Ruth E. Appel, PhD Candidate, Stanford University, rappel@stanford.edu. Jennifer Pan, Professor, Department of Communication, Stanford University, jp1@stanford.edu. Margaret E. Roberts, Professor, University of California, San Diego, meroberts@ucsd.edu. We thank Yanna Krupnikov and Eitan Hersh, Jon Roozenbeek, Lisa Opolka, Yannis Theocharis, Pablo Barberá, Andy Guess, and participants of the NYU CSMaP Conference, Stanford Polarization and Social Change Lab workshop for helpful comments; Antoni 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]. Ruth E. Appel has been supported by an SAP Stanford Graduate Fellowship in Science and Engineering, a Stanford Center on Philanthropy and Civil Society (PACS) PhD Research Fellowship, and a Stanford Impact Labs Summer Collaborative Research Fellowship.

1. Introduction

Misinformation is seen as a major global threat by political and economic leaders around the world (World Economic Forum 2022) as well as by the general public (Pew Research Center 2022; Silver 2022). 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 (Kozyreva et al. 2022). 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 (Bambauer 2021). 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 three weeks after its announcement (Lorenz 2022). 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.

In this paper, we seek to disaggregate the sources of partisan disagreement over what content social media companies should remove from the Internet. Drawing on the large literature on partisanship's influence on opinions toward public policies (Bartels 2016; Krupnikov et al. 2006; Lupia et al. 2007), we theorize that partisan differences in content moderation could stem from three different sources: 1) a "fact gap" – differences in what is perceived as misinformation; 2) a "value gap" – differences in overall preferences about how much misinformation should be removed; and 3) "party promotion" – a desire to leave misinformation online that promotes one's own party by flattering it or denigrating the other party.

We disaggregate the effects of the value gap and party promotion, holding the fact gap

constant, by embedding an experiment in a national survey of U.S. respondents where we present consenting participants misinformation headlines that they are explicitly told are false and where we vary the partisan alignment of the misinformation headline. We then ask respondents whether the social media company should remove the content, whether removal constitutes censorship, and whether they would report the content as harmful.

We find substantial support for the value gap. When Republicans and Democrats agree that the content is false, Republicans and Democrats still hold vastly different preferences for whether that information should be removed. Regardless of the partisan slant of the content, Democrats are more likely to support the removal of content, while Republicans are more likely to oppose removing content. We find evidence of the fact gap, with both Democrats and Republicans being more likely to believe that false headlines aligned with their own party are accurate. We find support for party promotion among Democrats who are slightly less likely to support removing misinformation that promotes their own side than misinformation that promotes Republicans, but no evidence of party promotion among Republicans. Finally, we find that party promotion among Democrats is reduced when accounting for beliefs about accuracy, though not completely explained by it, suggesting that the effect of party promotion on content moderation preferences of Democrats is overall quite small.

These results suggest that the value gap is an important mechanism underlying partisan preferences over content moderation of misinformation. The results are encouraging in that blatant partisan animosity does not seem to be the main driver of partisan disagreement over content moderation. Resolving the gap between Republicans' and Democrats' diverging values about content removal may be difficult, but identifying that there exists such difference paves the way for future work to shed light on what fears and concerns lie at the heart of this divergence and how this divergence could be addressed.

2. Sources of Partisan Disagreement

We theorize that partisan differences in content moderation could stem from three different sources: 1) a "fact gap" – differences in what is perceived as misinformation; 2) a "value gap" – differences in overall preferences about how much misinformation should be removed; and 3) "party promotion" – a desire to leave misinformation online that promotes one's own party or denigrates the other party, while removing misinformation that denigrates one's one party or promotes the other party.

In the first mechanism, the fact gap, partisanship influences how individuals reason about the information they encounter and its factual accuracy. A number of studies have found that there is disagreement over facts based on partisan identity (Bullock and Lenz 2019; Prior et al. 2015). Studies have also found that partisanship is correlated with how likely individuals are to believe misinformation, and decreases the likelihood that people can identify content as misinformation when the content aligns with their own political views and ideology (Allcott and Gentzkow 2017; Batailler et al. 2022; Flynn, Nyhan and Reifler 2017; Jakesch et al. 2019; Traberg and van der Linden 2022; Rhodes 2022; Roozenbeek et al. 2022; Gawronski 2021). This phenomenon may be due in part to related psychological mechanisms to preserve complementary beliefs (Kunda 1990; Lord, Ross and Lepper 1979; Taber and Lodge 2006), such as motivated reasoning, prior attitude effect, and confirmation bias, or to preserve one's identity (Ashokkumar et al. 2020; Van Bavel and Pereira 2018).[1] Others have argued that partisan differences in misinformation belief are driven by inattention (Pennycook and Rand 2019; Pennycook et al. 2021). When it comes to content removal, this suggests that even if partisans agree that misinformation should be removed (Knight Foundation and Ipsos 2022), the fact gap can lead to different preferences over what information should be removed because partisans disagree about what is true versus false.

While the fact gap has been the focus of many research efforts around the spread of misinformation (Allcott and Gentzkow 2017; Batailler et al. 2022; Flynn, Nyhan and Reifler 2017;

[1] See Ecker et al. (2022) for a review of the psychological drivers of misinformation beliefs.

Jakesch et al. 2019; Traberg and van der Linden 2022; Roozenbeek et al. 2022; Gawronski 2021), disagreement over facts may not be the only source of partisan divide over content removal. Even if partisans agree that a piece of content is false, they may disagree about whether it should be removed based on differences in values. Partisanship and support for specific political causes are tightly connected to identity, values, and personality (Campbell et al. 1980; Goren 2005; Ashokkumar et al. 2020; Van Bavel and Pereira 2018) and are often seen as deeply rooted (Green, Palmquist and Schickler 2002). As a result, those who identify as Democrat and Republican may simply have different preferences toward whether content, including misinformation, should ever be removed, with one side believing that the benefit of removing content outweighs the risk, and the other side believing that the risk outweighs the benefit. In other words, partisanship may influence individuals' reasoning about the need for and threats to free speech and expression, leading to a value gap. This partisan difference could be a result of external factors such as elite signaling (Ellis and Stimson 2012; Fiorina, Abrams and Pope 2005; McCarty, Poole and Rosenthal 2016) or internal factors such as personality, cognitive processes, motives, and emotions (Duckitt and Sibley 2010; Graham, Haidt and Nosek 2009; Jost, Federico and Napier 2009). Regardless, the value gap can result in partisan disagreement over content removal, even in the absence of a fact gap.

Lastly, even without a fact gap or value gap, partisans may disagree about the content to remove due to a desire to promote their own party. This party promotion mechanism might arise when individuals want to promote their own party by leaving misinformation online that benefits their own party or denigrates the other party, while removing misinformation that denigrates their side or promotes the other side. Partisanship might influence the desire to promote one's own party (in-group) relative to the other party (out-group), potentially due to the importance of the symbolic social standing of the in-group vs. the out-group (Huddy 2013). In the United States, party promotion may be a plausible potential mechanism given that affective polarization – the gap in affect toward the partisan in-group and the partisan out-group – has increased (Iyengar and Westwood 2015; Iyengar et al. 2019). Indeed,

studies have found evidence of party promotion in the US in other settings. For example, partisans will knowingly answer survey questions inaccurately to show support for their own side (for a review, see Bullock and Lenz (2019)). Several studies suggest that content flagging is sometimes used strategically to promote ones' own political aims rather than due to genuine belief (Crawford and Gillespie 2016). In an environment of intense partisan animus, partisans with the same views on whether content is false and the same preferences for whether misinformation should be removed may disagree on what content should be removed if they are trying to preserve misinformation that benefits their party and to remove misinformation that is unfavorable to their party.

3. Research Design

We embedded an experiment in a survey of U.S. adults commissioned by the Knight Foundation and fielded by Ipsos in the summer of 2021.[2] For our analysis, we focus on English-speaking respondents who identified as Democrat or Republican, resulting in 1,120 respondents, with mean age 53.29 ($SD = 16.53$) and 56.3% female (see Supplemental Information (SI) Descriptive Statistics section for detailed Descriptive Statistics). The experiment and analyses were pre-registered (see OSF project repository).[3] This research was approved by the Institutional Review Boards at Stanford University and the University of California, San Diego. Figure 1 shows an overview of the experiment design.

3.1. Treatment

Each participant was shown two different false news headlines sequentially (Headline 1 and Headline 2 in Figure 1). Respondents were told that "Someone has shared the following

[2]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 SI Descriptive Statistics section.

[3]We note deviations and clarification from the pre-analysis plan in the SI section Deviations, Clarifications and Additional Analyses and throughout the supplementary information text where they pertain. Replication code and data will be uploaded to the Harvard-MIT Dataverse once all analyses are finalized upon paper acceptance.

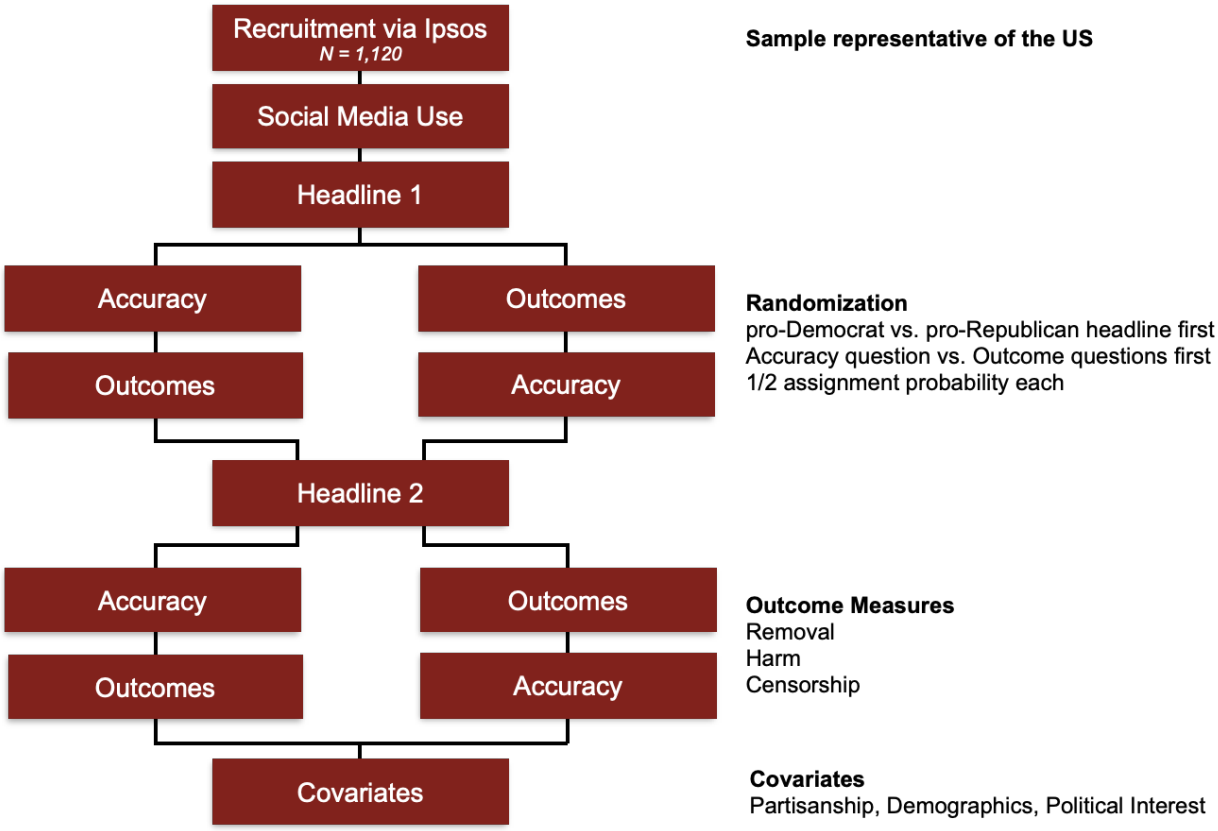


FIGURE 1. EXPERIMENT DESIGN OVERVIEW

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. These headlines were selected to cover similar topics that were similar in tone and pre-tested for partisan alignment in the expected direction and to ensure balance on other characteristics such as the intensity of the information conveyed (see SI section Headlines for additional details on selected headlines).

3.2. Randomization

This survey experiment relied on simple randomization at the participant and at the headline level. As mentioned above, participants were shown 2 different news headlines sequentially, randomly selected from a bank of 18 news headlines that contain false claims (see SI section Headlines). The news headlines were balanced in terms of partisanship, which implies random variation in which headlines were aligned or misaligned with a participant's partisanship. The order in which participants saw the headlines (i.e., pro-Democrat vs. pro-Republican first) was also randomized.

Sampling

This survey experiment was part of a larger survey commissioned by the Knight Foundation and implemented with the Ipsos KnowledgePanel and thus followed their general methods of including participants. Survey duration and sample size were limited in the context of this larger survey and we chose the maximum sample size available to us. Specifically, we

followed the exclusion criteria laid out in our pre-analysis plan. We only included participants who indicated that their preferred language is English. We excluded participants who are Independents because the alignment treatment would not work for Independents. We also excluded participants who had missing values for partisanship or indicated that they favored a party other than Democrats, Republicans, or Independents. Further, 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.

3.3. Measurement

All measures are described in detail in the SI section Questionnaire and Measures. Here, we provide a brief overview of our key measures.

Outcome Variables: The analysis focused on three main outcomes: (1) *Intent to Remove Headline (Removal)*: Whether or not the participant states that the headline should be removed by the social media company, (2) *Perception of Headline Removal as Censorship (Censorship)*: Whether the participant considers the removal of the headline censorship, (3) *Intent to Report Headline as Harmful (Harm)*: Whether the participant would report the headline as harmful content on a social media platform. 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 "Don't know" as a missing value. We deviated from the pre-analysis plan in recoding "Don't know" as a missing value instead of 0 because recoding "Don't 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 SI (see section A2.1.2), and find that the main results remain the same.

Accuracy: Even though we informed participants the headlines were rated as false by third-party fact checkers, they may have disagreed with this assessment. Therefore, we

asked respondents for the perceived accuracy of the false news headlines on a 4-point scale. Since we wanted to examine the effect of partisanship independent of assessments of accuracy, it was important that we measured participants' perceived accuracy of the false headlines.

To ensure our results were not driven by an "accuracy nudge" (Pennycook and Rand 2019; Pennycook et al. 2021), we randomized whether participants first answered the question about accuracy or the outcome questions after the headline (see Figure 1). 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 a perceived accuracy question after the outcome variables of Removal and Censorship were measured.[4]

Control variables and Indices: We measured a range of other variables such as news consumption and demographics as detailed in the SI section Questionnaire and Measures.[5] The order of response options in several questions on control variables, 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 the SI section Questionnaire and Measures.

3.4. Analysis

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

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

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 that respondent i is a Republican. The difference

[4]See SI Section A1.3.2 for balance tables, SI Section A2.1.3 for analysis with the first headline only, and SI Tables A67, A68, and A69 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.
[5]Controls include age, gender, education, race, ethnicity, household income, political interest, whether social media was the most common news source, whether a participant's posts had ever been flagged or they had been removed from social media.

in coefficients on D_i and R_i reflects the value gap, or the amount overall that Democrats and Republicans disagree about whether false content should be removed. Hd_a indicates that headline a is aligned for Democrats and Hr_a indicates that 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.

While we told all respondents that each headline is false, there were some respondents who still believed that the content is accurate. Therefore, to ensure that we have completely controlled for the fact gap, we conducted three additional analyses. First, we ran the same analyses with the just the subset of respondents who perceived the headline to be inaccurate. Second, we ran the same analysis only on headlines that on average both Republicans and Democrats think are most inaccurate and where there is little difference in perceived accuracy between Republicans and Democrats. This analysis helps us asses whether the results are driven by a particular headline. Third, we ran a mediation analysis of the effect of alignment with the headline on the outcomes, mediated by accuracy. Looking at the average direct effect of the mediation analysis allows us to better understand the direct effect of alignment, outside of the potentially mediating influence of accuracy.

4. Results

We find a large and statistically significant difference between the content moderation preferences of Republicans and Democrats. Overall, the probability Democrats say the false headline should be removed is 0.69, while the probability that Republicans say the false headline should be removed is 0.34. The probability Democrats would report a false headline as harmful is 0.49 while for Republicans it is 0.27. The probability that Democrats perceive removal of false headlines as censorship is 0.29 while for Republicans its 0.65.[6] Figure 2 plots the coefficient estimates and confidence interval (CI) from Equation 1 for all

[6]See Tables A55, A56, and A57 for regressions that calculate these probabilities.

respondents. The difference in estimates seen in the first row (Democrats) and second row (Republicans) of Figure 2 reveals a value gap between partisans when it comes to whether the false headlines should be removed from social media platforms (Panel A), whether the respondents would report a false headline as harmful (Panel B), and whether removal constitutes censorship (Panel C). This indicates large overall differences in content moderation preferences, regardless of headline alignment with political party.

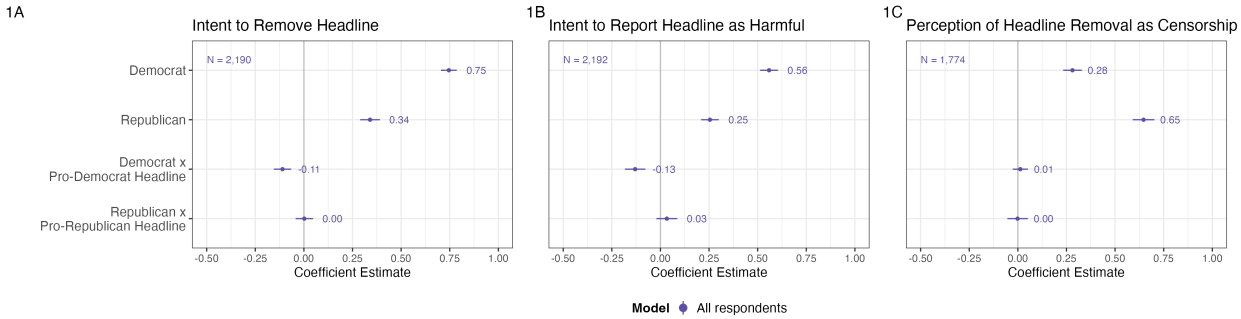


FIGURE 2. PARTISANSHIP AND PREFERENCES FOR CONTENT MODERATION FOR ALL RESPONDENTS (MODELS WITHOUT CONTROL VARIABLES).

The third row of coefficients in Figure 2 (estimate associated with Democrat x Pro-Democrat Headline) shows that Democrats are less likely to want to remove false headlines that promote the Democratic Party (Panel A, decrease in probability of 0.11) and less likely to report such content as harmful (Panel B, decrease in probability of 0.13). These results are statistically significant. In contrast, the fourth row of coefficients in Figure 2 (estimates associated with Republican x Pro-Republican Headline) shows that there are no statistically significant or meaningful differences among Republicans evaluating pro-Republican versus pro-Democrat headlines for whether they think the headline should be removed (Panel A), would report headlines as harmful (Panel B), or perceive removal as censorship (Panel C).

Perceptions of Accuracy: While we inform respondents that the headlines have been rated as false by third-party fact checkers, our results could be driven by respondents who do not believe that the headlines are false, regardless of our prompt. In this case, the fact gap, rather than a value gap or party promotion, might explain these differences. Indeed,

respondents rated 20.32% of headlines as either "Very accurate" or "Somewhat accurate" despite being told that they are false. Moreover, consistent with previous literature, the interaction terms in Figure 3 show that evaluations of the accuracy of the headline is partisan – both Democrats and Republicans are more likely to think that headlines that align with their own position are true, reflecting a fact gap. From Figure 3,[7] we see that Democrats rate 11% of Pro-Republican headlines as accurate but rate 25% of pro-Democrat headlines as accurate (11% + 14%). Similarly, Republicans rate 21% of pro-Democrat headlines as accurate and 32% of pro-Republican headlines as accurate (21% + 11%).

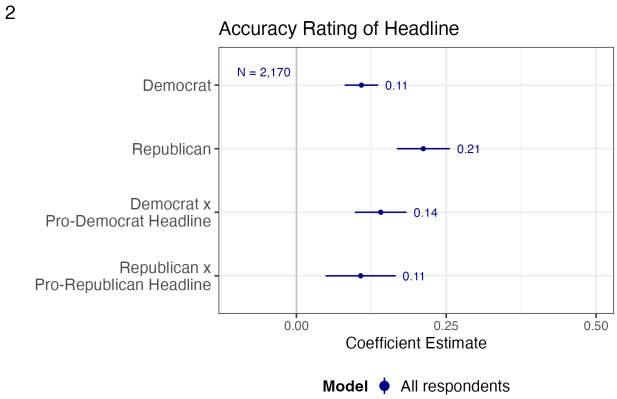


FIGURE 3. RESPONDENTS' ASSESSMENT OF HEADLINE ACCURACY (MODELS WITHOUT CONTROL VARIABLES). MODEL DOES NOT INCLUDE AN INTERCEPT TERM.

Inaccurate Subgroup Analysis: To further control for accuracy and the fact gap as a potential explanation, we first subset only to respondents who rated the headlines as inaccurate.[8] As Figure 4 shows, when we subset to respondents who agree that the false headlines are inaccurate, we see that the value 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. However, 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

[7]This analysis was not pre-registered.
[8]This includes respondents who assessed headlines as "Not very accurate" or "Not at all accurate".

Democrats is 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.

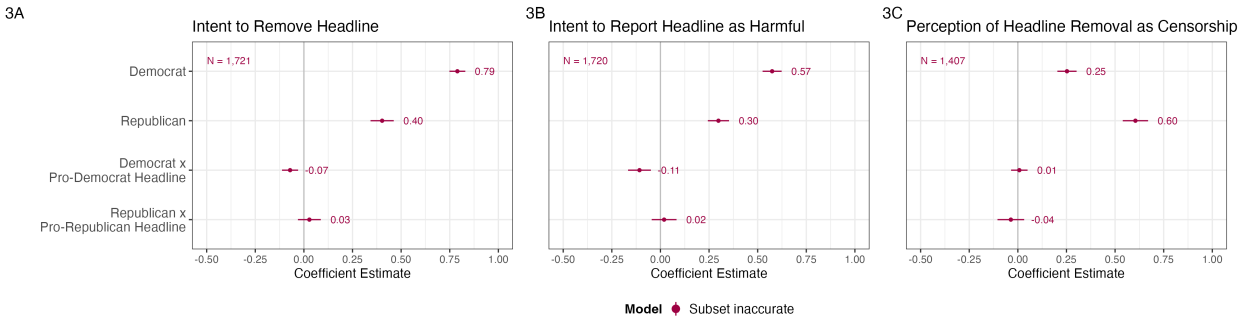


FIGURE 4. PARTISANSHIP AND PREFERENCES FOR CONTENT MODERATION FOR RESPONDENTS WHO AGREE THAT HEADLINES ARE INACCURATE (MODELS WITHOUT CONTROL VARIABLES).

Consensus Headlines Analysis: To push this further and to ensure that the evidence of a value gap in Figure 4 are not driven by any particular headline, we conduct these same analyses for 8 headlines that on average both Republicans and Democrats think are most inaccurate and where there is little difference in accuracy perception between Democrats and Republicans (see Tables in SI A2.1.5).[9] The reason we add this analysis is to address the concern that the gap in support for removal observed among Democrats and Republicans in the inaccurate subgroup analysis is driven by headlines with a large gap in perceived accuracy between Democrats and Republicans.[10] We continue to observe evidence for the value gap when we hone in on headlines Democrats and Republicans agree are inaccurate. Democrats remains 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.

[9]This analysis was not pre-registered.
[10]We also present results disaggregated headlines in SI A2.1.6. This analysis was not pre-registered.

Accuracy as Mediator: How much is party promotion among Democrats mediated by belief in the accuracy of the content? To study this, we conducted a mediation analysis for Democrat respondents where the effect of alignment of the respondent's partisanship with the headline's partisanship on each of the outcomes was mediated by the respondents' beliefs in accuracy of the content. Table 1 shows the results of this analysis. The Average Causal Mediation Effect (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, we observed party promotion on the intent to remove a headline and intent to report a headline as harmful for Democrats.[11] In Table 1, we see that the ACME is negative and significant for both intent to remove the headline and intent to report the headline as harmful, and the ADE is not significant for intent to remove the headline but is significant for intent to report the headline as harmful. This indicates that, while the party promotion effect for Democrats is significantly reduced when accounting for perceptions of accuracy, the fact gap cannot completely explain away party promotion among Democrats on all outcomes.

Previous research has found that prompting respondents with a question about accuracy may increase their attention to accuracy (Pennycook and Rand 2019; Pennycook et al. 2021). Given that we find that Democrats who rated the headlines as inaccurate were less likely to exhibit party promotion, we investigated whether an accuracy nudge might reduce party promotion. In Tables A67, A68, and A69 of the SI, we interact the party promotion effect with an indicator of whether accuracy was asked first.[12] While the estimates on these triple interactions are almost all positive, which would suggest that an accuracy nudge could be

[11] Table 1 includes Perception of Headline Removal as Censorship for completeness, but we do not observe party promotion in this outcome. As the total effect is not significant for this outcome, the ACME is difficult to interpret meaningfully.

[12] This analysis includes only the first headline that participants rated, because when participants rated the second headline, they had already seen the accuracy question regardless of whether they were randomized to see the accuracy question before or after outcomes, implying that accuracy had already been primed for all groups. This analysis was not pre-registered.

Table 1. Effect of Alignment Mediated by Accuracy for Democrats

Measure	Estimate	p-value
Intent to Remove Headline		
ACME	−0.065	< 0.001
ADE	−0.039	0.130
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	0.320	< 0.001
N Observations	1301	
N Simulations	1000	
Perception of Headline Removal as Censorship		
ACME	0.022	< 0.001
ADE	−0.015	0.584
Total Effect	0.006	0.852
Proportion Mediated	0.394	0.852
N Observations	1032	
N Simulations	1000	

Note: Mediation models were run with standard standard errors and without weigthing observations using a dataset in which missing values were addressed using listwise deletion. The models shown do not include any control variables.

used to reduce party promotion, none of them are significant on any outcome. We believe additional research in this area is may be warranted to see whether the accuracy nudge or a similar treatment could be used to alleviate the party promotion effect.

5. Discussion

The results of this experiment show that the value gap plays a huge role in attitudes toward removal of misinformation online. Republicans prefer that misinformation remains online and perceive removal of misinformation as censorship, even when they agree that the content is inaccurate. In contrast, Democrats strongly prefer that misinformation is removed and generally do not perceive such removal as censorship.

Why do we observe this value gap? The partisan value gap could be driven by differences in underlying fundamental principles that shape partisanship. Prior research finds that Republicans tend to emphasize freedom, purity, and individualistic values, while Democrats

value care and equality (Feinberg and Willer 2015; Silver and van Kessel 2021). These differences in underlying fundamental principles could explain why Democrats are more inclined to perceive misinformation as harmful and remove it, and why Republicans might want to protect freedom of expression and the purity of content as it was produced by users. This result may also stem from greater tolerance among Republicans for false statements (De keersmaecker and Roets 2019).

An alternative explanation for the value gap might be elite signaling. Previous research has shown that elite signaling can drive opinions (Ellis and Stimson 2012; Fiorina, Abrams and Pope 2005; McCarty, Poole and Rosenthal 2016). Republican elites have signaled their opposition to deplatforming and extensive content moderation, while Democrats have supported it. Thus, Republicans, knowing that their party is opposed to content removal, and Democrats, knowing that their party supports content moderation, might have selected responses that align with their respective party's positions. However, when Republican elites talk about online content removal, they are not signaling opposition to content removal per se. They are framing online content removal as a free speech and censorship issue,[13] that is, an issue pertaining to values that are important to Republicans. Indeed, we find that Republicans are more likely to perceive content removal as censorship than Democrats. Together, this suggests that if elite messaging and signaling plays a role, it is in framing and linking online content moderation to the core value of freedom of speech.

Further research is needed to explore how these findings generalize from a survey experiment to social media platforms. 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 (Rao, Morstatter and Lerman 2022). It could be that this difference in the prevalence of misinformation drives differences in the content moderation preferences of Democrats and

[13]For examples, see https://www.rubio.senate.gov/public/index.cfm/2021/6/rubio-introduces-sec-230-legislation-to-crack-down-on-big-tech-algorithms-and-protect-free-speech and https://www.flgov.com/2021/05/24/governor-ron-desantis-signs-bill-to-stop-the-censorship-of-floridians-by-big-tech/

Republicans. Alternatively, if Republican's threshold for unfollowing users mimics their high threshold for removing content, our findings could also explain why conservatives are exposed to more misinformation in general (Grinberg et al. 2019; Mosleh and Rand 2022; Nikolov, Flammini and Menczer 2020).

Beyond differences in prevalence, the content (e.g., political vs. health misinformation) and context (e.g., motivation to seek out the truth, how rooted beliefs on a topic are in one's identity) of misinformation headlines also matter (Knight Foundation and Ipsos 2022; Kozyreva et al. 2022; Ashokkumar et al. 2020; Pennycook and Rand 2022; Van Bavel and Pereira 2018). Future research should investigate further how specific types of content – non-political misinformation, hate speech, voter suppression content – and different contexts influence the value gap and party promotion. 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 value 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 (Douek 2022), 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 value gap might be less pronounced than for specific content. Second, future research could explore whether there might be partisan consensus on less extreme forms of content moderation, like flagging or down-weighting 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 (Feinberg and Willer 2019), to bridge the value gap.

Importantly, policymakers and social media platforms should understand that differences between Democrats and Republicans stem from more fundamental roots than disagreement over what is true versus false and partisan animosity. Instead, Americans seem to have diverging views on the principle of content removal and whether protection of free speech necessitates or precludes the moderation of content.

References

Allcott, Hunt and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31(2):211–236.
URL: *https://doi.org/10.1257/jep.31.2.211 https://www.aeaweb.org/content/file?id=4222*

Ashokkumar, Ashwini, Sanaz Talaifar, William T. Fraser, Rodrigo Landabur, Michael Buhrmester, Ángel Gómez, Borja Paredes and William B. Swann. 2020. "Censoring political opposition online: Who does it and why." *Journal of Experimental Social Psychology* 91(July).

Bambauer, Derek E. 2021. "What does the day after Section 230 reform look like?".
URL: *https://www.brookings.edu/techstream/what-does-the-day-after-section-230-reform-look-like/*

Bartels, Larry M. 2016. *Unequal Democracy*. Princeton, NJ: Princeton University Press.

Batailler, Cédric, Skylar M. Brannon, Paul E. Teas and Bertram Gawronski. 2022. "A Signal Detection Approach to Understanding the Identification of Fake News." *Perspectives on Psychological Science* 17(1):78–98.

Bullock, John G and Gabriel Lenz. 2019. "Partisan bias in surveys." *Annual Review of Political Science* 22:325–342.

Campbell, Angus, Philip E Converse, Warren E Miller and Donald E Stokes. 1980. *The American Voter*. University of Chicago Press.

Crawford, Kate and Tarleton Gillespie. 2016. "What is a flag for? Social media reporting tools and the vocabulary of complaint." *New Media & Society* 18(3):410–428.

De keersmaecker, Jonas and Arne Roets. 2019. "Is there an ideological asymmetry in the moral approval of spreading misinformation by politicians?" *Personality and Individual Differences* 143(January):165–169.
URL: *https://doi.org/10.1016/j.paid.2019.02.003*

Douek, Evelyn. 2022. "Content Moderation as Systems Thinking." *Harvard Law Review* 136.

Duckitt, John and Chris G Sibley. 2010. "Personality, ideology, prejudice, and politics: A dual-process motivational model." *Journal of personality* 78(6):1861–1894.

Ecker, Ullrich K. H., Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga and Michelle A. Amazeen. 2022. "The psychological drivers of misinformation belief and its resistance to correction." *Nature Reviews Psychology* 1(1):13–29.

Ellis, Christopher and James A Stimson. 2012. *Ideology in America*. Cambridge University Press.

Feinberg, Matthew and Robb Willer. 2015. "From Gulf to Bridge: When Do Moral

Arguments Facilitate Political Influence?" *Personality and Social Psychology Bulletin* 41(12):1665–1681.

Feinberg, Matthew and Robb Willer. 2019. "Moral reframing: A technique for effective and persuasive communication across political divides." *Social and Personality Psychology Compass* 13(12):1–12.

Fiorina, Morris P, Samuel J Abrams and Jeremy C Pope. 2005. *Culture war: The myth of a polarized America*. Pearson Longman.

Flynn, DJ, Brendan Nyhan and Jason Reifler. 2017. "The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics." *Political Psychology* 38:127–150.

Gawronski, Bertram. 2021. "Partisan bias in the identification of fake news." *Trends in Cognitive Sciences* 25(9):723–724.

Goren, Paul. 2005. "Party identification and core political values." *American Journal of Political Science* 49(4):881–896.

Graham, Jesse, Jonathan Haidt and Brian A Nosek. 2009. "Liberals and conservatives rely on different sets of moral foundations." *Journal of personality and social psychology* 96(5):1029.

Green, Donald, Bradley Palmquist and Eric Schickler. 2002. *Partisan hearts and minds: Political parties and the social identities of voters*. New Haven, CT: Yale University Press.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson and David Lazer. 2019. "Fake news on Twitter during the 2016 US presidential election." *Science* 363(6425):374–378.

Huddy, Leonie. 2013. From Group Identity to Political Cohesion and Commitment. In *The Oxford Handbook of Political Psychology*, ed. Leonie Huddy, David O. Sears and Jack S. Levy. 2nd ed. Oxford University Press pp. 737–773.

Iyengar, Shanto and Sean J Westwood. 2015. "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59(3):690–707.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." *Annual Review of Political Science* 22(1):129–146.

Jakesch, Maurice, Moran Koren, Anna Evtushenko and Mor Naaman. 2019. "The Role of Source and Expressive Responding in Political News Evaluation." wording: aligned vs. unaligned.

URL: *https://www.mauricejakesch.com/assets/pdf/cpj2019_source_headline_news.pdf*

Jost, John T, Christopher M Federico and Jaime L Napier. 2009. "Political ideology: Its structure, functions, and elective affinities." *Annual review of psychology* 60:307–337.

Knight Foundation and Ipsos. 2022. Free Expression in America Post-2020. Technical report.
URL: *https://knightfoundation.org/wp-content/uploads/2022/01/KF_Free_Expression_2022.pdf*

Kozyreva, Anastasia, Stefan Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser and Jason Reifler. 2022. "Free speech vs. harmful misinformation: Moral dilemmas in online content moderation." Published on PsyArXiv.

Krupnikov, Yanna, Adam Seth Levine, Arthur Lupia and Markus Prior. 2006. "Public ignorance and estate tax repeal: The effect of partisan differences and survey incentives." *National Tax Journal* 59(3):425–437.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3):480–498.

Lord, Charles G., Lee Ross and Mark R. Lepper. 1979. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." *Journal of Personality and Social Psychology* 37(11):2098–2109.

Lorenz, Taylor. 2022. "How the Biden administration let right-wing attacks derail its disinformation efforts.".
URL: *https://www.washingtonpost.com/technology/2022/05/18/disinformation-board-dhs-nina-jankowicz/*

Lupia, Arthur, Adam Seth Levine, Jesse O. Menning and Gisela Sin. 2007. "Were bush tax cut supporters "simply ignorant?" A second look at conservatives and liberals in "homer gets a tax cut"." *Perspectives on Politics* 5(4):773–784.

McCarty, Nolan, Keith T Poole and Howard Rosenthal. 2016. *Polarized America: The dance of ideology and unequal riches*. mit Press.

Mosleh, Mohsen and David G Rand. 2022. "Measuring exposure to misinformation from political elites on Twitter." *Nature Communications* 13(1):1–9.

Nikolov, Dimitar, Alessandro Flammini and Filippo Menczer. 2020. "Right and left, partisanship predicts (asymmetric) vulnerability to misinformation." *arXiv preprint arXiv:2010.01462* .

Pennycook, Gordon and David G. Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188(June 2018):39–50.
URL: *https://doi.org/10.1016/j.cognition.2018.06.011*

Pennycook, Gordon and David G. Rand. 2022. "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation." *Nature Communications* 13(1).

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio Alonso Arechar, Dean Eckles and David Gertler Rand. 2021. "Shifting attention to accuracy can reduce misinformation

online." *Nature* 592:590–595.

Pew Research Center. 2022. Climate Change Remains Top Global Threat Across 19-Country Survey. Technical report.
URL: *https://www.pewresearch.org/global/wp-content/uploads/sites/2/2022/08/PG_2022.08.31_Global-Threats_FINAL.pdf*

Prior, Markus, Gaurav Sood, Kabir Khanna et al. 2015. "You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions." *Quarterly Journal of Political Science* 10(4):489–518.

Rao, Ashwin, Fred Morstatter and Kristina Lerman. 2022. "Partisan Asymmetries in Exposure to Misinformation." *Scientific Reports* 12(15671):1–11.
URL: *http://arxiv.org/abs/2203.01350*

Rhodes, Samuel C. 2022. "Filter bubbles, echo chambers, and fake news: how social media conditions individuals to be less critical of political misinformation." *Political Communication* 39(1):1–22.

Roozenbeek, Jon, Rakoen Maertens, Stefan M Herzog, Michael Geers, Ralf Kurvers, ¶ Mubashir and Sander Van Der Linden. 2022. "Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking Susceptibility to misinformation." *Judgment and Decision Making* 17(3):547–573.
URL: *http://orcid.org/0000-0002-8150-9305*

Silver, Laura. 2022. "Americans see different global threats facing the country now than in March 2020.".
URL: *https://www.pewresearch.org/fact-tank/2022/06/06/americans-see-different-global-threats-facing-the-country-now-than-in-march-2020/*

Silver, Laura and Patrick van Kessel. 2021. Both Republicans and Democrats prioritize family, but they differ over other sources of meaning in life. Technical report Pew Research Center.
URL: *https://www.pewresearch.org/fact-tank/2021/11/22/both-republicans-and-democrats-prioritize-family-but-they-differ-over-other-sources-of-meaning-in-life/*

Taber, Charles S and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3):755–769.

Traberg, Cecilie Steenbuch and Sander van der Linden. 2022. "Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility." *Personality and Individual Differences* 185(September 2021):111269.
URL: *https://doi.org/10.1016/j.paid.2021.111269*

Van Bavel, Jay J. and Andrea Pereira. 2018. "The Partisan Brain: An Identity-Based Model of Political Belief." *Trends in Cognitive Sciences* 22(3):213–224.
URL: *http://dx.doi.org/10.1016/j.tics.2018.01.004*

World Economic Forum. 2022. The Global Risks Report 2022. Technical report.
URL: *https://www.weforum.org/reports/global-risks-report-2022*

Supplemental Information

Table of Contents

A1. Extended Materials and Methods

A1.1. Deviations, Clarifications and Additional Analyses

Deviation #1: In the PAP, we wrote that all outcome measures are binary with the exception of the censorship measure, which was recoded as a binary measure by considering "Yes " as 1, and "No" and "Don't know" as 0. We code "Don't know" as a missing value instead of 0 because recoding "Don't 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 Main Models With All Headlines and Pre-Registered Censorship Coding section, and find that the main results remain the same.

Deviation #2: In the PAP, we said we would estimate models for those who perceived the headline to be accurate, and separately for those who perceived the headline to be inaccurate. We deviated from this by presenting models for all participants vs. the subgroup of participants that perceived the headline as inaccurate. Since models with all participants contain both the accurate and the inaccurate subgroup and is therefore an average of the two sub-groups, the difference between all observations and the inaccurate subgroup gives us insight into the accurate subgroup.

Deviation #3: For the models including controls, we deviate from the PAP and did not constrain β_D and β_R to be the same. This is because different than in the case of Allcott and Gentzkow (2017), the effects for Democrats and Republicans were not similar and we did not want to risk masking true differences by analyzing a combined effect.

Deviation #4: The selection of control variables was constrained by the data that Ipsos collected. Some of the control variables we had asked for – political efficacy, affective polarization, voting behavior in 2020, political news consumption, whether a user ever used social media, whether a user had ever been banned from social media – were not implemented in the survey and hence do not appear as control variables in our analysis.

Deviation #5: One variable we had requested measuring how much participants think that political information from a range of different sources, including print media and social media, can be trusted, was not included as requested. The dataset did include a grid of trust questions that measured whether participants trusted social media companies, the news media, that the news media reported in an unbiased manner, and institutions like the government. However, those variables did not focus on political news, and were only assigned to half of respondents, perhaps because it was part of another experiment in the larger Ipsos survey. Because of the different nature of the question and the low number of responses, we could not include any control variable for trust in media.

Clarification #1: We remove participants for whom no survey language was indicated and participants that indicated proficiency in Spanish only.

Clarification #2: We excluded participants who had missing values for partisanship or indicated that they favored a party other than Democrats, Republicans, or Independents.

Clarification #3: We removed 243 participants who were part of a student sample that was different from the sample meant to be representative of the U.S. population.

Additional analyses: We ran additional analyses that were not pre-registered: regressions of the main outcomes considering only the first headline that participants rated, regressions

without interaction effects, regressions with consensus headlines only, regressions disaggregated by headline, and regressions including a triple interaction between accuracy question order, participant partisanship, and headline alignment; a regression of perceived headline accuracy on partisanship and alignment.

A1.2. Analysis

A1.2.1. Modeling

Similar to the approach used by Allcott and Gentzkow (2017), we ran regression analyses with interaction terms for partisanship of participants and political alignment of the headlines:

$$censorship_{ia} = \beta_D D_i \cdot Hd_a + \beta_R R_i \cdot Hr_a + \gamma_D \cdot D_i + \gamma_R R_i + \varepsilon_{ia} \tag{A1}$$

$$removal_{ia} = \beta_D D_i \cdot Hd_a + \beta_R R_i \cdot Hr_a + \gamma_D \cdot D_i + \gamma_R R_i + \varepsilon_{ia} \tag{A2}$$

$$harm_{ia} = \beta_D D_i \cdot Hd_a + \beta_R R_i \cdot Hr_a + \gamma_D \cdot D_i + \gamma_R R_i + \varepsilon_{ia} \tag{A3}$$

$censorship_{ia}$ is a binary measure of whether an individual i rated the removal of headline a as censorship.

$removal_{ia}$ is a binary measure of whether an individual i thinks the social media platform should remove the headline a from its platform.

$harm_{ia}$ is a binary measure of whether an individual i would report the content of the headline a as harmful to the social media platform.

D_i indicates that respondent i is a Democrat.

R_i indicates that respondent i is a Republican.

Hd_a is an indicator that headline a is aligned with Democratic views.

Hr_a is an indicator that headline a is aligned with Republican views.

β_D measures whether a Democrat is more likely to (1) perceive removal of pro-Democratic content as censorship (in equation A1), (2) think the social media platform should remove a pro-Democratic headline (in equation A2), or (3) report the content of a pro-Democratic headline as harmful to the social media platform (in equation A3).

β_R measures whether a Republican is more likely to (1) perceive removal of pro-Republican content as censorship (in equation A1), (2) think the social media platform should remove a pro-Republican headline (in equation A2), or (3) report the content of a pro-Republican headline as harmful to the social media platform (in equation A3).

Given that the headlines are balanced in terms of political alignment and randomly assigned to participants, the estimated β parameters measure the effect of political alignment.

In the pre-analysis plan, we said we would control for perceived accuracy by using a sub-group analysis to estimate the models just for those who perceived the headline to be accurate, and separately for those who perceived the headline to be inaccurate. We deviated in that we present models for all participants vs. the sub-group of participants that perceived

the headline as inaccurate. This still allows us to evaluate whether the partisanship of the headline influences evaluations of whether it should be removed and whether removal would be considered censorship, among those who evaluated the accuracy of the headline in the same way because all observations contain both the accurate and the inaccurate subgroup, and are therefore an average of the two sub-groups, so a difference between all observations and the inaccurate subgroup implies a difference between the two subgroups. Additionally, our main interest is in analyzing the views of the inaccurate subgroup, allowing us to evaluate how participants reacted to misinformation headlines they believed were false.

We first ran all specifications first without control variables. We also ran specifications with controls (see Questionnaire and Measures for a list of control variables).

For the models including controls, deviating from the pre-analysis plan, we did not constrain β_D and β_R to be the same, because different than in the case of Allcott and Gentzkow (2017), the effects for Democrats and Republicans were not similar and we did not want to risk masking true differences by analyzing a combined effect.

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 the Additional Results section.

In addition to the pre-registered analyses, we ran regressions of our main outcomes considering only the first headline that participants rated, regressions without interaction effects, regressions with consensus headlines only, regressions disaggregated by headline, and regressions including a triple interaction between accuracy question order, participant partisanship, and headline alignment, and a regression of perceived headline accuracy on partisanship and alignment.. We show all results that are not already shown in the main text in the Additional Results section.

A1.3. Data

A1.3.1. Missing Data

For variables with missing data, we (1) imputed missing data using the `Amelia` package in R (Honaker, King and Blackwell 2011), and (2) used listwise deletion to remove observations with missing data. We show the results for both approaches in the Additional Results section.

A1.3.2. Balance Checks

Here, we present balance tables of control variables across the different experiment arms (aligned vs. misaligned headlines, accuracy questions displayed before vs. after treatment).

Overall, the different experimental groups are relatively balanced. For partisan alignment (see Table A1), the Hispanic indicator has the highest Standardized Mean Difference (SMD), but the randomization seems to have been effective. For the accuracy question order (see Table A2), the control variables education, household income and whether social media is the most common news format have relatively high SMD, suggesting that it is worthwhile to include these control variables in some of our models.

TABLE A1. BALANCE TABLE FOR PARTISAN ALIGNMENT OF HEADLINE, FIRST HEADLINE

Variable	Aligned		p-value	SMD
	Yes	No		
Number of Observations	558	562		
Age (mean (SD))	53.74 (16.40)	52.84 (16.67)	0.360	0.055
Gender = Female (N (%))	307 (55.0)	324 (57.7)	0.408	0.053
Education (N (%))			0.948	0.051
... No high school diploma or GED	24 (4.3)	25 (4.4)		
... High school graduate	148 (26.5)	143 (25.4)		
... Some college or Associate degree	168 (30.1)	179 (31.9)		
... Bachelor's degree	128 (22.9)	121 (21.5)		
... Master's degree or above	90 (16.1)	94 (16.7)		
Hispanic = Yes (N (%))	85 (15.2)	63 (11.2)	0.057	0.119
Race = Non-White (N (%))	235 (42.1)	256 (45.6)	0.272	0.069
Household Income (N (%))			0.769	0.109
... Under $10,000	16 (2.9)	12 (2.1)		
... $10,000 to $24,999	41 (7.3)	41 (7.3)		
... $25,000 to $49,999	99 (17.7)	90 (16.0)		
... $50,000 to $74,999	101 (18.1)	97 (17.3)		
... $75,000 to $99,999	83 (14.9)	83 (14.8)		
... $100,000 to $149,999	97 (17.4)	119 (21.2)		
... $150,000 or more	121 (21.7)	120 (21.4)		
Political Interest (mean (SD))	2.82 (0.68)	2.82 (0.67)	0.936	0.005
Social Media Most Common News Format = Yes (N (%))	81 (14.8)	95 (17.2)	0.309	0.066
Social Media Post Flagged = Yes (N (%))	63 (14.2)	69 (15.6)	0.617	0.040
Social Media Post Removed = Yes (N (%))	57 (12.6)	62 (13.9)	0.637	0.038

Note: p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. Standardized mean difference (SMD) and p-values are exactly the same for the subset of data on the second headline, only the data in the Yes and No columns would be reversed, therefore we show only one table.

TABLE A2. BALANCE TABLE FOR ACCURACY QUESTION ORDER

Variable	Accuracy Question Order		p-value	SMD
	First	Second		
Number of Observations	581	539		
Age (mean (SD))	53.00 (16.28)	53.59 (16.81)	0.551	0.036
Gender = Female (N (%))	331 (57.0)	300 (55.7)	0.702	0.026
Education (N (%))			0.048	0.186
... No high school diploma or GED	23 (4.0)	26 (4.8)		
... High school graduate	133 (22.9)	158 (29.3)		
... Some college or Associate degree	200 (34.4)	147 (27.3)		
... Bachelor's degree	128 (22.0)	121 (22.4)		
... Master's degree or above	97 (16.7)	87 (16.1)		
Hispanic = Yes (N (%))	77 (13.3)	71 (13.2)	1.000	0.002
Race = Non-White (N (%))	255 (43.9)	236 (43.8)	1.000	0.002
Household Income (N (%))			0.089	0.199
... Under $10,000	14 (2.4)	14 (2.6)		
... $10,000 to $24,999	34 (5.9)	48 (8.9)		
... $25,000 to $49,999	100 (17.2)	89 (16.5)		
... $50,000 to $74,999	100 (17.2)	98 (18.2)		
... $75,000 to $99,999	75 (12.9)	91 (16.9)		
... $100,000 to $149,999	125 (21.5)	91 (16.9)		
... $150,000 or more	133 (22.9)	108 (20.0)		
Political Interest (mean (SD))	2.82 (0.69)	2.83 (0.66)	0.773	0.017
Social Media Most Common News Format = Yes (N (%))	77 (13.5)	99 (18.6)	0.024	0.141
Social Media Post Flagged = Yes (N (%))	62 (13.2)	70 (16.7)	0.182	0.096
Social Media Post Removed = Yes (N (%))	66 (14.1)	53 (12.4)	0.510	0.051

Note: p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. Accuracy question order was randomized at the participant level, therefore balance checks were run on the short data frame with one headline observations per participant.

A1.4. Headlines

Below is the bank of false news headlines. These false news headlines are based on headlines from Snopes.com, a major established fact checking site. On May 19, 2021, we went through the 50 most recent pages of the Snopes Politics Archive (https://www.snopes.com/fact-check/category/politics/) to retrieve recent fact-checked fake news headlines.

The fact checks on these pages were published between May 18, 2021, and December 22, 2020, in the Politics category. The recency of the headlines ensures that they are not outdated and come from the same political period. We only considered claims that were labelled as "false" (i.e., we excluded "mixed", "mostly false", "true", and any other category). Additionally, because there were far fewer fake news headlines that were aligned for Democrats, we looked beyond the initial dates and identified a false claim on Snopes from an earlier date that still seemed relevant and was studied in a recent conference paper (Mosleh et al. 2021). We then excluded headlines that did not have a clear partisan slant,[A1] were miscategorized (e.g., related to the Business category on Snopes), or were outdated. Finally, out of all headlines, we selected headlines aligned for either Democrats or Republicans in pairs, where each pair was relatively comparable in terms of the intensity of the false information (e.g., headlines involving physical violence – murder, torture, mutilation – would have greater intensity than those that deal with incompetence) and the topic they cover (e.g., racism, protest).

To maximize ecological validity, we formatted headlines in ways similar to how headlines would appear on social media. We created a template headline in the Facebook format, and then took the claims from Snopes (or the original news headline, if the primary source was a news article, cited in Snopes) as the headline text.[A2] We then appended an image, either from the related Snopes article, the primary source, or a search for images related to the headline text via Google.

We pretested the headlines for partisan alignment and excluded one headline that was perceived as neutral rather than aligned for either party, as well as another headline that was not perceived as strongly aligned in order to rebalance the number of headlines aligned for either party. For all other headlines in the headline bank, the pretest with a convenience sample of $N = 20$ showed partisan alignment in the expected direction. The headlines were also relatively balanced in terms of the extent to which they were aligned for each part and in terms of their perceived intensity (i.e., how worrying or exciting they seemed to participants).

[A1] This includes headlines with less well-known political figures or headlines that required additional context to easily understand them.

[A2] Some headlines have slight modifications from what is shown on Snopes or the original, e.g., removing punctuation at the end of a claim.

Pro-Democrat	*Pro-Republican*

Donald Trump clones White House with his replica Oval Office at Florida home

Pro-Democrat 1

Hours after signing an executive order on Jan. 20, 2021, U.S. President Joe Biden violated his own mask mandate

Pro-Republican 1

85% of Americans approved of U.S. President Joe Biden's first speech before a joint session of Congress

Pro-Democrat 2

Biden warns if Americans don't get COVID jabs they might have to cancel July 4

Pro-Republican 2

In Sept. 2016, Ted Cruz tweeted, "I'll believe in climate change when Texas freezes over."

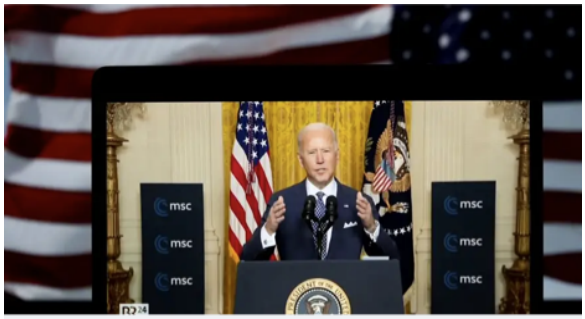
Pro-Democrat 3

How Biden's climate plan could limit you to eat just one burger a MONTH

Pro-Republican 3

"If English was good enough for Jesus, it's good enough for us," stated Marjorie Taylor Greene, explaining why Spanish shouldn't be spoken in a Christian Nation like America

Pro-Democrat 4



On Feb. 19, 2021, U.S. President Joe Biden uttered a racial slur referring to Black people in a recorded speech

Pro-Republican 4

Former U.S. first lady Melania Trump had her own private bedroom on Air Force One

Pro-Democrat 5

Smoking Gun! Joe Biden's Dept. of Energy Blocked Texas from Increasing Power Ahead of Killer Storm

Pro-Republican 5

Former U.S. President Donald Trump left his successor, President Joe Biden, a letter that read, "Joe, you know I won."

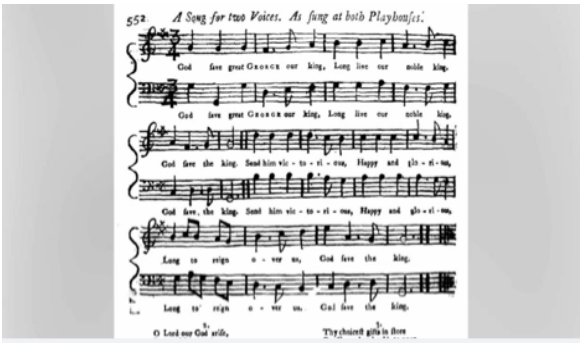
Pro-Democrat 6

U.S. President Joe Biden's lead secret service agent is Chinese

Pro-Republican 6

Donald Trump once evicted a disabled combat veteran for owning a small therapy dog

PRO-DEMOCRAT 7



Oxford wants to scrap sheet music over complicity in 'white supremacy'

PRO-REPUBLICAN 7

Former U.S. Vice President Mike Pence unfollowed President Donald Trump following Jan. 6, 2021, riots in Washington, D.C.

PRO-DEMOCRAT 8

Democrats are trying to change Mother's Day to Birthing People's Day

PRO-REPUBLICAN 8

U.S. Sen. Ted Cruz messed up when he attributed the famous phrase "full of sound and fury / signifying nothing" to William Shakespeare

PRO-DEMOCRAT 9

Rep. Alexandria Ocasio-Cortez said "I hope I never have twins because I can't afford to be pregnant for 18 months."

PRO-REPUBLICAN 9

A1.5. Questionnaire and Measures

Randomization and branching logic instructions provided to Ipsos are shown in cyan.
Choice display and configuration instructions provided to Ipsos are shown in magenta.
Numeric coding or transformation for variables as in the final dataset and related information about the inclusion of variables is shown in green, surrounded by brackets.
Variable names as they appear in the regression tables and plots are shown in violet, but were not shown to participants.
Headers are displayed in bold and grey and are for orientation only and was not displayed to participants.
Question numbering was not displayed to participants.

Survey experiment measures are listed first and control variables second. We had asked Ipsos to measure the control first (i.e., pre-treatment). Additionally, for some questions, the exact question wording was not provided since they were part of the standard demographics or sample variables.

A1.5.1. Survey Experiment Measures

1. **Headline display**
   We have a bank of 9 pro-Democrat (headline_pro_dem_[number].png) and 9 pro-Republican (headline_pro_rep_[number].png) headline images. Each respondent should be shown 1 pro-Democrat image and 1 pro-Republican image. Whether the respondent see a pro-Democrat or pro-Republican image first should be randomized

2. **Headline 1**

   1. Someone has shared the following headline on a social media site. (This headline has been established as **false** by third party fact checkers.)
      Randomly select 1 (pro_rep/pro_dem) image and Insert image file here
      The image should remain visible for the censorship, accuracy, and harm questions that follow

3. **Measuring outcomes 1**
   Randomize question order of Censorship 1 and Accuracy 1: for half of respondents, the two "Censorship 1" questions should be shown first, for the other half, the "Accuracy 1" question should be shown first

4. **Censorship 1**

   1. Intent to Remove Headline: How do you think the social media company should handle this headline?
      Single answer allowed

1. Allow it to remain on the social media site [0]
2. Remove it from the social media site [1]

2. Perception of Headline Removal as Censorship: Imagine that the social media platform removed this headline. Would you use the word "censorship" to describe this action?
Single answer allowed
   1. Yes [1]
   2. No [0]
   3. Don't know [NA]

5. **Accuracy 1**

1. Accuracy: To the best of your knowledge, how accurate is the claim in the above headline?
Single answer allowed
   1. Not at all accurate [1]
   2. Not very accurate [2]
   3. Somewhat accurate [3]
   4. Very accurate [4]

6. **Harm 1**

1. Intent to Report Headline as Harmful: Some social media platforms allow users to report content as harmful. If you have the option of anonymously reporting this content as harmful, would you click the "report as harmful content" button for the above headline?
Single answer allowed
   1. Yes [1]
   2. No [0]

7. **Headline 2**

1. Someone has shared the following headline on a social media site. (This headline has been established as **false** by third party fact checkers.)
Randomly select 1 (pro_rep/pro_dem) image and Insert image file here
The image should remain visible for the censorship, accuracy, and harm questions that follow

8. **Measuring outcomes 2**
Display censorship and accuracy questions for headline 2 in the same order as those for headline 1: If participants saw "Censorship 1" before "Accuracy 1", they should see "Censorship 2" before "Accuracy2"; if participants saw "Accuracy 1" before "Censorship 1", they should see "Accuracy 2" before "Censorship 2"

9. **Censorship 2**

    1. Intent to Remove Headline: How do you think the social media company should handle this headline?
    Single answer allowed

        1. Allow it to remain on the social media site [0]
        2. Remove it from the social media site [1]

    2. Perception of Headline Removal as Censorship: Imagine that the social media platform removed this headline. Would you use the word "censorship" to describe this action?
    Single answer allowed

        1. Yes [1]
        2. No [0]
        3. Don't know [NA]

10. **Accuracy 2**

    1. Accuracy: To the best of your knowledge, how accurate is the claim in the above headline?
    Single answer allowed

        1. Not at all accurate [1]
        2. Not very accurate [2]
        3. Somewhat accurate [3]
        4. Very accurate [4]

11. **Harm 2**

    1. Intent to Report Headline as Harmful: Some social media platforms allow users to report content as harmful. If you have the option of anonymously reporting this content as harmful, would you click the "report as harmful content" button for the above headline?
    Single answer allowed

        1. Yes [1]
        2. No [0]

A1.5.2. Control Variables

The selection of control variables was constrained by the data that Ipsos collected. Some of the variables we had asked for – political efficacy, affective polarization, voting behavior in 2020, political news consumption, whether a user ever used social media, whether a user had ever been banned from social media – were not implemented in the survey and hence do not appear in the final survey data here. Other variables, like partisanship and political interest, were worded differently from those in our pre-analysis plan. Some variables that

we had not originally requested, such as whether a participant's social media posts had been flagged in the past or what their most common news source was, were used as proxies for variables that were not provided. As mentioned in our pre-analysis plan, we rely on the measures that Ipsos actually provided. The demographic variables were not listed explicitly in the pre-analysis plan because they are part of the general demographic information about a sample that Ipsos provides and we mentioned we assumed these will already be included. Here, we provide the final version of all variables that we use for analysis.
One variable we had requested measuring how much participants think that political information from a range of different sources, including print media and social media, can be trusted, was not included as requested. The dataset did include a grid of trust questions that measured whether participants trusted social media companies, the news media, that the news media reported in an unbiased manner, and institutions like the government. However, those variables did not focus on political news, and the questions related to trust in media seemed to be part of another experiment in the larger Ipsos survey because only half of participants answered the question on trust in the news media, while the other half answered the question on the unbiasedness of news media reporting. Because of the different nature of the question and the low number of participants having responded to it, we did not include any control variable for trust in media.

13. **Partisanship**

    1. Generally speaking, do you think of yourself as...
       Select one answer only.
       Single answer allowed
        1. Republican: Republican [Republican]
        2. Democrat: Democrat [Democrat]
        3. Independent [NA; excluded from anlysis]
        4. Something else [NA; excluded from analysis]

14. **Social media use**

    1. Social Media Post Removed: Have you ever experienced the following? - Had a social media post removed by the social media company.
       Single answer allowed
        1. Yes, I have experienced or done this [1]
        2. No, I have not experienced or done this [0]
        3. Not applicable [NA]
    2. Social Media Post Flagged: Have you ever experienced the following? - Had a social media post flagged, reported, or tagged with a warning label.
       Single answer allowed [We had originally requested a variable asking whether a participant was ever banned from a social media platform, which was not included in the final survey. We therefore included this variable as a proxy for past experiences with social media content moderation.]

1. Yes, I have experienced or done this [1]
2. No, I have not experienced or done this [0]
3. Not applicable [NA]

3. Social Media Most Common News Format: In which format do you get most of your news?
Single answer allowed
[We had originally requested a variable asking whether a participant ever uses social media. This variable was not included in the final survey. Instead, we included another measure of social media use that we had not originally requested, but was part of the data: whether social media was a participant's most common news source, which was derived recoding a variable asking respondents for the most common news source.]

1. From a printed newspaper or magazine [0]
2. From television [0]
3. From radio [0]
4. From social media [1]
5. From friends and family [0]

15. **Political interest**

1. Political Interest: How closely do you follow each of these different news topics?
Grid: News about national issues and politics, News about your state government, News about issues affecting your local community, International affairs
Single answer allowed for each source in the grid
[We had originally requested a single variable asking participants how often they pay attention to what is going on in government and politics. Since Ipsos provided a grid of related variables, we calculated and index calculated as average of values on these variables. We performed a factor analysis and found that all individual variables load onto the same factor.]

1. Very closely [4]
2. Somewhat closely [3]
3. Not too closely [2]
4. Not at all closely [1]

16. **Demographics**

1. Age: Age
Single answer allowed
2. Gender: Gender
Single answer allowed

1. Male [0]
2. Female [1]

3. Education: Education
Single answer allowed
   1. No high school diploma or GED [1]
   2. High school graduate (high school diploma or the equivalent GED) [2]
   3. Some college or Associate degree [3]
   4. Bachelor's degree [4]
   5. Master's degree or above [5]

4. Household Income: Household Income
Single answer allowed
   1. Under $10,000 [1]
   2. $10,000 to $24,999 [2]
   3. $25,000 to $49,999 [3]
   4. $50,000 to $74,999 [4]
   5. $75,000 to $99,999 [5]
   6. $100,000 to $149,999 [6]
   7. $150,000 or more [7]

5. Hispanic: Hispanic Origin
Single answer allowed
[Mexican/Mexican-American/Chicano; Puerto Rican; Cuban, Cuban-American; Other Spanish/Hispanic/Latino were recoded as Yes]
   1. Yes [1]
   2. No [0]

6. Race: Race
Single answer allowed
[Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian/Pacific Islander, 2+ races were recoded as Non-White]
   1. White [0]
   2. Non-White [1]

A1.5.3. Derived Variables and Variables Based on Stimuli

17. Pro-Democrat Headline: Indicates whether a headline is pro-Democrat, either by making Democrats look good or by making Republicans look bad. [1 if true, 0 otherwise]

18. Pro-Republican Headline: Indicates whether a headline is pro-Republican, either by making Republicans look good or by making Democrats look bad. [1 if true, 0 otherwise]

19. Aligned: Indicates whether a participant's partisanship and headline orientation are aligned (i.e., Democrat partisanship and pro-Democrat headline, or Republican partisanship and pro-Republican headline). [1 if true, 0 otherwise]

20. Accuracy Binary: Divides participants into two subgroups for each headline they see, one subgroup that considers the misinformation headline as accurate, one that considers the misinformation headline as inaccurate. [1 if the rating on Accuracy was "Somewhat accurate" or "Very accurate", 0 if rating on Accuracy was "Not at all accurate" or "Not very accurate"]

21. Accuracy Order: Order in which Accuracy question appeared. [1 if Accuracy questions came first (before Censorship outcome questions), 0 if Accuracy questions came second (after Censorship outcome questions)]

A1.6. Descriptive Statistics

TABLE A3. DESCRIPTIVE STATISTICS

Variable	N	Mean	SD	Min	Q1	Median	Q3	Max
Age	1120	53.288	16.534	18	40	55	66	94
Gender	1120							
... Male	489	43.7%						
... Female	631	56.3%						
Education	1120							
... No high school diploma or GED	49	4.4%						
... High school graduate	291	26%						
... Some college or Associate degree	347	31%						
... Bachelor's degree	249	22.2%						
... Master's degree or above	184	16.4%						
Hispanic	1120							
... Yes	148	13.2%						
... No	972	86.8%						
Race	1120							
... White	629	56.2%						
... Non-White	491	43.8%						
Household Income	1120							
... Under $10,000	28	2.5%						
... $10,000 to $24,999	82	7.3%						
... $25,000 to $49,999	189	16.9%						
... $50,000 to $74,999	198	17.7%						
... $75,000 to $99,999	166	14.8%						
... $100,000 to $149,999	216	19.3%						
... $150,000 or more	241	21.5%						
Political Interest	1101	2.822	0.675	1	2.5	3	3.25	4
Social Media Most Common News Format	1102							
... Yes	176	16%						
... No	926	84%						
Social Media Post Flagged	888							
... Yes	132	14.9%						
... No	756	85.1%						
Social Media Post Removed	896							
... Yes	119	13.3%						
... No	777	86.7%						
Partisanship	1120							
... Democrat	673	60.1%						
... Republican	447	39.9%						

Table A4. Descriptive Statistics on Representativeness of Sample

Sample	Weighting	Median Age	Share Hispanic or Latino
US Population	unweighted	38.2	0.187
Iposos full sample (including students)	unweighted	48.0	0.174
Iposos full sample (including students)	weighted	40.5	0.166
Iposos full sample (excluding students)	unweighted	54.0	0.171
Iposos full sample (excluding students)	weighted	48.5	0.166
Iposos sample for this study (excluding students)	unweighted	54.0	0.132
Iposos sample for this study (excluding students)	weighted	49.5	0.126
Final sample	unweighted	55.0	0.132
Final sample	weighted	50.5	0.123

Sources: For 2020 age data: U.S. Census Bureau, 2016-2020 American Community Survey 5-Year Estimates, retrieved on September 7, 2022 from https://data.census.gov/cedsci/table?q=median%20age&g=0100000US&tid=ACSST5Y2020.S0101. For 2020 ethnicity data: U.S. Census Bureau, 2020 Census Redistricting Data (Public Law 94-171), retrieved on September 7, 2022 from https://data.census.gov/cedsci/table?q=hispanic&g=0100000US&tid=DECENNIALPL2020.P2.

A2. Additional Results

A2.1. Regression Tables

We show results in seven sections: The first section shows the main regression models. The second section shows robustness checks for the main regression models where the censorship outcome is coded as pre-registered and not as mentioned in the Deviations, Clarifications and Additional Analyses section. The third section shows a robustness check for the main regression models including only the first headline that participants rated in the regressions. The fourth section shows regressions similar to the main regressions, but without interaction effects. The fifth section shows the main regression models when restricting the headlines to the consensus headlines only. The sixth section shows the regression results when disaggregating the models by headline. The seventh section shows models with a triple interaction between accuracy question order, participant partisanship, and headline alignment. As mentioned in the Deviations, Clarifications and Additional Analyses section, the analyses shown in the third through the seventh section were not pre-registered.

A2.1.1. Main Models Considering All Headlines

Intent to Remove Headline

Results Based on Listwise Deletion

Weighted Model Using Clustered SEs

Table A5. Regression of Intent to Remove Headline on Partisanship and Alignment

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.75***	0.79***	0.79***	0.98***
	(0.02)	(0.12)	(0.02)	(0.13)
Republican	0.34***	0.37**	0.40***	0.60***
	(0.03)	(0.12)	(0.03)	(0.14)
Democrat x Pro-Democrat Headline	−0.11***	−0.12***	−0.07***	−0.07**
	(0.02)	(0.03)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.00	0.01	0.03	0.03
	(0.02)	(0.03)	(0.03)	(0.03)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.01		0.00
		(0.03)		(0.04)
Education		−0.01		−0.04·
		(0.02)		(0.02)
Hispanic		0.07		0.09·
		(0.05)		(0.05)
Race: Non-White		−0.01		0.03
		(0.04)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.02		−0.02
		(0.03)		(0.03)
Social Media Most Common News Format		0.03		0.01
		(0.05)		(0.06)
Social Media Post Flagged		0.01		0.03
		(0.07)		(0.09)
Social Media Post Removed		−0.15*		−0.21*
		(0.07)		(0.09)
R^2	0.58	0.59	0.65	0.67
Adj. R^2	0.58	0.59	0.65	0.67
Num. obs.	2190	1691	1721	1349
RMSE	0.46	0.46	0.45	0.44
N Clusters	1104	849	1003	783

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A6. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.76***	0.75***	0.80***	0.93***
	(0.02)	(0.10)	(0.02)	(0.11)
Republican	0.35***	0.34***	0.40***	0.53***
	(0.02)	(0.10)	(0.03)	(0.11)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.06***	−0.06**
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.00	0.01	0.03	0.03
	(0.02)	(0.02)	(0.03)	(0.03)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.02		0.02
		(0.03)		(0.03)
Education		0.00		−0.02
		(0.01)		(0.02)
Hispanic		0.08*		0.09·
		(0.04)		(0.04)
Race: Non-White		0.00		0.03
		(0.03)		(0.03)
Household Income		−0.01		−0.01
		(0.01)		(0.01)
Political Interest		−0.02		−0.02
		(0.02)		(0.03)
Social Media Most Common News Format		0.02		0.02
		(0.04)		(0.04)
Social Media Post Flagged		0.04		0.04
		(0.05)		(0.06)
Social Media Post Removed		−0.19***		−0.22***
		(0.05)		(0.06)
R^2	0.62	0.63	0.69	0.70
Adj. R^2	0.62	0.62	0.69	0.70
Num. obs.	2190	1691	1721	1349
RMSE	0.46	0.46	0.45	0.44
N Clusters	1104	849	1003	783

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A7. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.75***	0.79***	0.79***	0.98***
	(0.02)	(0.10)	(0.02)	(0.11)
Republican	0.34***	0.37***	0.40***	0.60***
	(0.03)	(0.10)	(0.03)	(0.11)
Democrat x Pro-Democrat Headline	−0.11***	−0.12***	−0.07*	−0.07*
	(0.03)	(0.04)	(0.03)	(0.04)
Republican x Pro-Republican Headline	0.00	0.01	0.03	0.03
	(0.04)	(0.04)	(0.04)	(0.05)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.01		0.00
		(0.03)		(0.03)
Education		−0.01		−0.04*
		(0.01)		(0.02)
Hispanic		0.07·		0.09*
		(0.04)		(0.04)
Race: Non-White		−0.01		0.03
		(0.03)		(0.03)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.02		−0.02
		(0.02)		(0.03)
Social Media Most Common News Format		0.03		0.01
		(0.04)		(0.04)
Social Media Post Flagged		0.01		0.03
		(0.05)		(0.07)
Social Media Post Removed		−0.15**		−0.21**
		(0.05)		(0.07)
R^2	0.58	0.59	0.65	0.67
Adj. R^2	0.58	0.59	0.65	0.67
Num. obs.	2190	1691	1721	1349
RMSE	0.46	0.46	0.45	0.44

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A8. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.76***	0.75***	0.80***	0.93***
	(0.02)	(0.08)	(0.02)	(0.09)
Republican	0.35***	0.34***	0.40***	0.53***
	(0.02)	(0.08)	(0.03)	(0.09)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.06*	−0.06*
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.00	0.01	0.03	0.03
	(0.03)	(0.04)	(0.04)	(0.04)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.02		0.02
		(0.02)		(0.02)
Education		0.00		−0.02·
		(0.01)		(0.01)
Hispanic		0.08*		0.09*
		(0.03)		(0.04)
Race: Non-White		0.00		0.03
		(0.03)		(0.03)
Household Income		−0.01		−0.01
		(0.01)		(0.01)
Political Interest		−0.02		−0.02
		(0.02)		(0.02)
Social Media Most Common News Format		0.02		0.02
		(0.03)		(0.04)
Social Media Post Flagged		0.04		0.04
		(0.04)		(0.05)
Social Media Post Removed		−0.19***		−0.22***
		(0.04)		(0.05)
R^2	0.62	0.63	0.69	0.70
Adj. R^2	0.62	0.62	0.69	0.70
Num. obs.	2190	1691	1721	1349
RMSE	0.46	0.46	0.45	0.44

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Results Based on Multiple Imputation

Weighted Model Using Clustered SEs

TABLE A9. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.74***	0.67***	0.79***	0.79***
	(0.02)	(0.11)	(0.02)	(0.12)
Republican	0.34***	0.28**	0.40***	0.44***
	(0.03)	(0.10)	(0.03)	(0.12)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.08***	−0.07**
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.00	0.00	0.03	0.02
	(0.02)	(0.02)	(0.03)	(0.03)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.03		0.03
		(0.03)		(0.03)
Education		−0.01		−0.03
		(0.02)		(0.02)
Hispanic		0.09*		0.12*
		(0.04)		(0.04)
Race: Non-White		0.01		0.06
		(0.03)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.00		0.01
		(0.02)		(0.03)
Social Media Most Common News Format		0.01		−0.00
		(0.04)		(0.05)
Social Media Post Flagged		−0.02		−0.01
		(0.05)		(0.06)
Social Media Post Removed		−0.11*		−0.15*
		(0.05)		(0.06)
R^2	0.58	0.59	0.65	0.66
Adj. R^2	0.58	0.59	0.65	0.65
Num. obs.	2240	2240	1782	1782
N Clusters	1120	1120	1035	1035

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A10. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.76***	0.65***	0.80***	0.77***
	(0.02)	(0.08)	(0.02)	(0.09)
Republican	0.35***	0.26**	0.40***	0.41***
	(0.02)	(0.08)	(0.03)	(0.10)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.06***	−0.06**
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.00	0.00	0.03	0.03
	(0.02)	(0.02)	(0.03)	(0.03)
Age		0.00·		0.00
		(0.00)		(0.00)
Gender: Female		0.05·		0.04
		(0.02)		(0.03)
Education		0.01		−0.01
		(0.01)		(0.01)
Hispanic		0.10**		0.11**
		(0.04)		(0.04)
Race: Non-White		0.02		0.06*
		(0.03)		(0.03)
Household Income		−0.01		−0.01
		(0.01)		(0.01)
Political Interest		−0.00		0.00
		(0.02)		(0.02)
Social Media Most Common News Format		−0.00		−0.01
		(0.04)		(0.04)
Social Media Post Flagged		−0.01		0.01
		(0.04)		(0.04)
Social Media Post Removed		−0.13**		−0.16**
		(0.04)		(0.05)
R^2	0.62	0.63	0.68	0.69
Adj. R^2	0.62	0.62	0.68	0.69
Num. obs.	2240	2240	1782	1782
N Clusters	1120	1120	1035	1035

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A11. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.74***	0.67***	0.79***	0.79***
	(0.02)	(0.09)	(0.02)	(0.10)
Republican	0.34***	0.28**	0.40***	0.44***
	(0.03)	(0.09)	(0.03)	(0.10)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.08*	−0.07*
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.00	0.00	0.03	0.02
	(0.04)	(0.04)	(0.04)	(0.04)
Age		0.00·		0.00
		(0.00)		(0.00)
Gender: Female		0.03		0.03
		(0.02)		(0.03)
Education		−0.01		−0.03·
		(0.01)		(0.01)
Hispanic		0.09**		0.12**
		(0.03)		(0.04)
Race: Non-White		0.01		0.06*
		(0.03)		(0.03)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.00		0.01
		(0.02)		(0.02)
Social Media Most Common News Format		0.01		−0.00
		(0.03)		(0.04)
Social Media Post Flagged		−0.02		−0.01
		(0.04)		(0.05)
Social Media Post Removed		−0.11**		−0.15**
		(0.04)		(0.05)
R^2	0.58	0.59	0.65	0.66
Adj. R^2	0.58	0.59	0.65	0.65
Num. obs.	2240	2240	1782	1782

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A12. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.76***	0.65***	0.80***	0.77***
	(0.02)	(0.07)	(0.02)	(0.08)
Republican	0.35***	0.26***	0.40***	0.41***
	(0.02)	(0.07)	(0.03)	(0.08)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.06*	−0.06*
	(0.03)	(0.02)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.00	0.00	0.03	0.03
	(0.03)	(0.03)	(0.04)	(0.04)
Age		0.00*		0.00
		(0.00)		(0.00)
Gender: Female		0.05*		0.04·
		(0.02)		(0.02)
Education		0.01		−0.01
		(0.01)		(0.01)
Hispanic		0.10***		0.11***
		(0.03)		(0.03)
Race: Non-White		0.02		0.06*
		(0.02)		(0.02)
Household Income		−0.01		−0.01
		(0.01)		(0.01)
Political Interest		−0.00		0.00
		(0.02)		(0.02)
Social Media Most Common News Format		−0.00		−0.01
		(0.03)		(0.03)
Social Media Post Flagged		−0.01		0.01
		(0.03)		(0.04)
Social Media Post Removed		−0.13***		−0.16***
		(0.04)		(0.05)
R^2	0.62	0.63	0.68	0.69
Adj. R^2	0.62	0.62	0.68	0.69
Num. obs.	2240	2240	1782	1782

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Intent to Report Headline as Harmful

Results Based on Listwise Deletion

Weighted Model Using Clustered SEs

Table A13. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.55***	0.57***	0.58***
	(0.02)	(0.10)	(0.02)	(0.12)
Republican	0.25***	0.26**	0.30***	0.32**
	(0.02)	(0.09)	(0.03)	(0.11)
Democrat x Pro-Democrat Headline	−0.13***	−0.11***	−0.11***	−0.08*
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.03	0.04	0.02	0.02
	(0.03)	(0.03)	(0.03)	(0.04)
Age		0.00**		0.00**
		(0.00)		(0.00)
Gender: Female		−0.06·		−0.05
		(0.03)		(0.04)
Education		−0.01		−0.02
		(0.02)		(0.02)
Hispanic		−0.01		0.04
		(0.05)		(0.05)
Race: Non-White		0.04		0.04
		(0.04)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.04		−0.02
		(0.02)		(0.03)
Social Media Most Common News Format		0.04		0.04
		(0.04)		(0.05)
Social Media Post Flagged		−0.03		−0.03
		(0.05)		(0.06)
Social Media Post Removed		−0.10·		−0.12·
		(0.05)		(0.07)
R^2	0.42	0.42	0.46	0.46
Adj. R^2	0.42	0.41	0.46	0.45
Num. obs.	2192	1692	1720	1347
RMSE	0.47	0.46	0.47	0.47
N Clusters	1105	851	1005	785

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

Table A14. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.55***	0.58***	0.56***
	(0.02)	(0.10)	(0.02)	(0.11)
Republican	0.27***	0.27**	0.31***	0.30**
	(0.02)	(0.09)	(0.02)	(0.11)
Democrat x Pro-Democrat Headline	−0.11***	−0.09***	−0.09***	−0.08**
	(0.02)	(0.02)	(0.02)	(0.03)
Republican x Pro-Republican Headline	0.03	0.02	0.03	0.01
	(0.02)	(0.02)	(0.03)	(0.03)
Age		0.00**		0.00**
		(0.00)		(0.00)
Gender: Female		−0.06*		−0.05
		(0.03)		(0.03)
Education		−0.01		−0.03
		(0.01)		(0.02)
Hispanic		−0.02		0.03
		(0.04)		(0.05)
Race: Non-White		0.03		0.04
		(0.03)		(0.04)
Household Income		0.00		−0.00
		(0.01)		(0.01)
Political Interest		−0.03		−0.02
		(0.02)		(0.03)
Social Media Most Common News Format		−0.00		−0.01
		(0.04)		(0.05)
Social Media Post Flagged		0.01		0.00
		(0.05)		(0.06)
Social Media Post Removed		−0.12*		−0.12·
		(0.05)		(0.07)
R^2	0.45	0.44	0.48	0.48
Adj. R^2	0.45	0.44	0.48	0.47
Num. obs.	2192	1692	1720	1347
RMSE	0.48	0.47	0.49	0.48
N Clusters	1105	851	1005	785

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A15. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.55***	0.57***	0.58***
	(0.02)	(0.10)	(0.02)	(0.11)
Republican	0.25***	0.26**	0.30***	0.32**
	(0.02)	(0.09)	(0.03)	(0.11)
Democrat x Pro-Democrat Headline	−0.13***	−0.11**	−0.11**	−0.08*
	(0.03)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	0.03	0.04	0.02	0.02
	(0.03)	(0.04)	(0.04)	(0.04)
Age		0.00***		0.00**
		(0.00)		(0.00)
Gender: Female		−0.06*		−0.05·
		(0.03)		(0.03)
Education		−0.01		−0.02
		(0.01)		(0.02)
Hispanic		−0.01		0.04
		(0.04)		(0.05)
Race: Non-White		0.04		0.04
		(0.03)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.04·		−0.02
		(0.02)		(0.03)
Social Media Most Common News Format		0.04		0.04
		(0.04)		(0.04)
Social Media Post Flagged		−0.03		−0.03
		(0.05)		(0.06)
Social Media Post Removed		−0.10*		−0.12*
		(0.05)		(0.06)
R^2	0.42	0.42	0.46	0.46
Adj. R^2	0.42	0.41	0.46	0.45
Num. obs.	2192	1692	1720	1347
RMSE	0.47	0.46	0.47	0.47

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A16. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.55***	0.58***	0.56***
	(0.02)	(0.08)	(0.02)	(0.10)
Republican	0.27***	0.27***	0.31***	0.30**
	(0.02)	(0.08)	(0.02)	(0.09)
Democrat x Pro-Democrat Headline	−0.11***	−0.09**	−0.09**	−0.08*
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.03	0.02	0.03	0.01
	(0.03)	(0.03)	(0.04)	(0.04)
Age		0.00***		0.00***
		(0.00)		(0.00)
Gender: Female		−0.06*		−0.05·
		(0.02)		(0.03)
Education		−0.01		−0.03·
		(0.01)		(0.01)
Hispanic		−0.02		0.03
		(0.04)		(0.04)
Race: Non-White		0.03		0.04
		(0.03)		(0.03)
Household Income		0.00		−0.00
		(0.01)		(0.01)
Political Interest		−0.03		−0.02
		(0.02)		(0.02)
Social Media Most Common News Format		−0.00		−0.01
		(0.03)		(0.04)
Social Media Post Flagged		0.01		0.00
		(0.04)		(0.05)
Social Media Post Removed		−0.12**		−0.12*
		(0.04)		(0.06)
R^2	0.45	0.44	0.48	0.48
Adj. R^2	0.45	0.44	0.48	0.47
Num. obs.	2192	1692	1720	1347
RMSE	0.48	0.47	0.49	0.48

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Results Based on Multiple Imputation

Weighted Model Using Clustered SEs

TABLE A17. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.52***	0.57***	0.53***
	(0.02)	(0.09)	(0.02)	(0.11)
Republican	0.26***	0.23**	0.30***	0.27*
	(0.02)	(0.08)	(0.03)	(0.10)
Democrat x Pro-Democrat Headline	−0.13***	−0.13***	−0.11***	−0.11***
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.03	0.03	0.02	0.01
	(0.03)	(0.03)	(0.03)	(0.03)
Age		0.00**		0.00**
		(0.00)		(0.00)
Gender: Female		−0.05·		−0.03
		(0.03)		(0.03)
Education		−0.01		−0.02
		(0.01)		(0.02)
Hispanic		0.01		0.06
		(0.04)		(0.05)
Race: Non-White		0.03		0.02
		(0.03)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.01		0.01
		(0.02)		(0.03)
Social Media Most Common News Format		0.04		0.05
		(0.04)		(0.05)
Social Media Post Flagged		−0.04		−0.03
		(0.04)		(0.06)
Social Media Post Removed		−0.07		−0.08
		(0.05)		(0.06)
R^2	0.43	0.43	0.46	0.47
Adj. R^2	0.42	0.43	0.46	0.46
Num. obs.	2240	2240	1782	1782
N Clusters	1120	1120	1035	1035

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

Table A18. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.50***	0.58***	0.51***
	(0.02)	(0.08)	(0.02)	(0.10)
Republican	0.27***	0.21**	0.31***	0.25**
	(0.02)	(0.08)	(0.02)	(0.10)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.10***	−0.10***
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.02	0.02	0.02	0.02
	(0.02)	(0.02)	(0.03)	(0.03)
Age		0.00***		0.00**
		(0.00)		(0.00)
Gender: Female		−0.04		−0.03
		(0.02)		(0.03)
Education		−0.01		−0.02
		(0.01)		(0.01)
Hispanic		−0.00		0.04
		(0.04)		(0.04)
Race: Non-White		0.03		0.03
		(0.03)		(0.03)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.00		0.01
		(0.02)		(0.02)
Social Media Most Common News Format		−0.00		−0.00
		(0.04)		(0.04)
Social Media Post Flagged		−0.02		−0.01
		(0.05)		(0.06)
Social Media Post Removed		−0.08		−0.08
		(0.05)		(0.06)
R^2	0.45	0.46	0.48	0.49
Adj. R^2	0.45	0.45	0.48	0.49
Num. obs.	2240	2240	1782	1782
N Clusters	1120	1120	1035	1035

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A19. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.52***	0.57***	0.53***
	(0.02)	(0.08)	(0.02)	(0.10)
Republican	0.26***	0.23**	0.30***	0.27**
	(0.02)	(0.08)	(0.03)	(0.10)
Democrat x Pro-Democrat Headline	−0.13***	−0.13***	−0.11**	−0.11**
	(0.03)	(0.03)	(0.04)	(0.04)
Republican x Pro-Republican Headline	0.03	0.03	0.02	0.01
	(0.03)	(0.03)	(0.04)	(0.04)
Age		0.00***		0.00**
		(0.00)		(0.00)
Gender: Female		−0.05*		−0.03
		(0.02)		(0.03)
Education		−0.01		−0.02
		(0.01)		(0.01)
Hispanic		0.01		0.06
		(0.04)		(0.04)
Race: Non-White		0.03		0.02
		(0.03)		(0.03)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.01		0.01
		(0.02)		(0.02)
Social Media Most Common News Format		0.04		0.05
		(0.04)		(0.04)
Social Media Post Flagged		−0.04		−0.03
		(0.04)		(0.05)
Social Media Post Removed		−0.07		−0.08
		(0.05)		(0.06)
R^2	0.43	0.43	0.46	0.47
Adj. R^2	0.42	0.43	0.46	0.46
Num. obs.	2240	2240	1782	1782

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A20. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.56***	0.50***	0.58***	0.51***
	(0.02)	(0.07)	(0.02)	(0.08)
Republican	0.27***	0.21**	0.31***	0.25**
	(0.02)	(0.07)	(0.02)	(0.08)
Democrat x Pro-Democrat Headline	−0.11***	−0.11***	−0.10**	−0.10**
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.02	0.02	0.02	0.02
	(0.03)	(0.03)	(0.04)	(0.04)
Age		0.00***		0.00***
		(0.00)		(0.00)
Gender: Female		−0.04·		−0.03
		(0.02)		(0.02)
Education		−0.01		−0.02
		(0.01)		(0.01)
Hispanic		−0.00		0.04
		(0.03)		(0.04)
Race: Non-White		0.03		0.03
		(0.02)		(0.03)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.00		0.01
		(0.02)		(0.02)
Social Media Most Common News Format		−0.00		−0.00
		(0.03)		(0.03)
Social Media Post Flagged		−0.02		−0.01
		(0.04)		(0.05)
Social Media Post Removed		−0.08·		−0.08
		(0.04)		(0.05)
R^2	0.45	0.46	0.48	0.49
Adj. R^2	0.45	0.45	0.48	0.49
Num. obs.	2240	2240	1782	1782

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Perception of Headline Removal as Censorship

Results Based on Listwise Deletion

Weighted Model Using Clustered SEs

Table A21. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.28***	0.12	0.25***	0.07
	(0.02)	(0.15)	(0.03)	(0.17)
Republican	0.65***	0.51**	0.60***	0.45**
	(0.03)	(0.14)	(0.03)	(0.16)
Democrat x Pro-Democrat Headline	0.01	0.03	0.01	0.03
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	−0.00	0.01	−0.04	−0.03
	(0.03)	(0.03)	(0.03)	(0.04)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		0.02
		(0.04)		(0.04)
Education		−0.00		0.01
		(0.02)		(0.02)
Hispanic		−0.10·		−0.06
		(0.06)		(0.06)
Race: Non-White		0.04		0.04
		(0.04)		(0.05)
Household Income		0.00		−0.00
		(0.01)		(0.01)
Political Interest		0.03		0.03
		(0.03)		(0.04)
Social Media Most Common News Format		−0.01		−0.01
		(0.06)		(0.06)
Social Media Post Flagged		−0.08		−0.11
		(0.06)		(0.08)
Social Media Post Removed		0.19**		0.22*
		(0.06)		(0.08)
R^2	0.53	0.55	0.47	0.49
Adj. R^2	0.53	0.55	0.47	0.48
Num. obs.	1774	1406	1407	1125
RMSE	0.46	0.46	0.45	0.45
N Clusters	958	753	851	677

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A22. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.27***	0.17	0.24***	0.13
	(0.02)	(0.11)	(0.02)	(0.13)
Republican	0.64***	0.56***	0.60***	0.50***
	(0.03)	(0.11)	(0.03)	(0.13)
Democrat x Pro-Democrat Headline	0.01	0.01	0.00	0.02
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.01	0.01	−0.03	−0.02
	(0.02)	(0.03)	(0.03)	(0.03)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.03		−0.01
		(0.03)		(0.03)
Education		−0.02		−0.01
		(0.02)		(0.02)
Hispanic		−0.08·		−0.06
		(0.05)		(0.05)
Race: Non-White		0.01		0.00
		(0.04)		(0.04)
Household Income		0.01		0.01
		(0.01)		(0.01)
Political Interest		0.03		0.03
		(0.03)		(0.03)
Social Media Most Common News Format		−0.04		−0.06
		(0.04)		(0.05)
Social Media Post Flagged		−0.06		−0.09
		(0.05)		(0.06)
Social Media Post Removed		0.18**		0.21**
		(0.06)		(0.07)
R^2	0.51	0.53	0.45	0.47
Adj. R^2	0.51	0.53	0.45	0.47
Num. obs.	1774	1406	1407	1125
RMSE	0.46	0.45	0.46	0.45
N Clusters	958	753	851	677

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A23. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.28***	0.12	0.25***	0.07
	(0.02)	(0.12)	(0.03)	(0.13)
Republican	0.65***	0.51***	0.60***	0.45***
	(0.03)	(0.11)	(0.03)	(0.13)
Democrat x Pro-Democrat Headline	0.01	0.03	0.01	0.03
	(0.03)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	−0.00	0.01	−0.04	−0.03
	(0.04)	(0.04)	(0.05)	(0.05)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		0.02
		(0.03)		(0.03)
Education		−0.00		0.01
		(0.02)		(0.02)
Hispanic		−0.10*		−0.06
		(0.04)		(0.05)
Race: Non-White		0.04		0.04
		(0.03)		(0.04)
Household Income		0.00		−0.00
		(0.01)		(0.01)
Political Interest		0.03		0.03
		(0.03)		(0.03)
Social Media Most Common News Format		−0.01		−0.01
		(0.04)		(0.05)
Social Media Post Flagged		−0.08		−0.11·
		(0.05)		(0.06)
Social Media Post Removed		0.19**		0.22**
		(0.06)		(0.07)
R^2	0.53	0.55	0.47	0.49
Adj. R^2	0.53	0.55	0.47	0.48
Num. obs.	1774	1406	1407	1125
RMSE	0.46	0.46	0.45	0.45

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A24. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.27***	0.17·	0.24***	0.13
	(0.02)	(0.09)	(0.02)	(0.10)
Republican	0.64***	0.56***	0.60***	0.50***
	(0.03)	(0.09)	(0.03)	(0.10)
Democrat x Pro-Democrat Headline	0.01	0.01	0.00	0.02
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.01	0.01	−0.03	−0.02
	(0.04)	(0.04)	(0.04)	(0.05)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.03		−0.01
		(0.02)		(0.03)
Education		−0.02		−0.01
		(0.01)		(0.01)
Hispanic		−0.08*		−0.06
		(0.04)		(0.04)
Race: Non-White		0.01		0.00
		(0.03)		(0.03)
Household Income		0.01		0.01
		(0.01)		(0.01)
Political Interest		0.03		0.03
		(0.02)		(0.02)
Social Media Most Common News Format		−0.04		−0.06
		(0.03)		(0.04)
Social Media Post Flagged		−0.06		−0.09·
		(0.04)		(0.05)
Social Media Post Removed		0.18***		0.21***
		(0.05)		(0.06)
R^2	0.51	0.53	0.45	0.47
Adj. R^2	0.51	0.53	0.45	0.47
Num. obs.	1774	1406	1407	1125
RMSE	0.46	0.45	0.46	0.45

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Results Based on Multiple Imputation

Weighted Model Using Clustered SEs

Table A25. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.30***	0.30*	0.27***	0.28·
	(0.02)	(0.12)	(0.02)	(0.14)
Republican	0.63***	0.62***	0.59***	0.58***
	(0.03)	(0.11)	(0.03)	(0.13)
Democrat x Pro-Democrat Headline	0.02	0.02	0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.00	0.00	−0.03	−0.03
	(0.03)	(0.03)	(0.04)	(0.04)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		0.01
		(0.03)		(0.03)
Education		0.00		0.01
		(0.02)		(0.02)
Hispanic		−0.11*		−0.08
		(0.05)		(0.05)
Race: Non-White		0.01		−0.01
		(0.03)		(0.04)
Household Income		−0.00		−0.00
		(0.01)		(0.01)
Political Interest		−0.00		−0.02
		(0.02)		(0.03)
Social Media Most Common News Format		0.00		0.00
		(0.05)		(0.06)
Social Media Post Flagged		−0.01		−0.00
		(0.06)		(0.07)
Social Media Post Removed		0.11·		0.11
		(0.06)		(0.07)
R^2	0.52	0.53	0.47	0.48
Adj. R^2	0.52	0.52	0.47	0.47
Num. obs.	2240	2240	1782	1782
N Clusters	1120	1120	1035	1035

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A26. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.28***	0.35***	0.26***	0.31**
	(0.02)	(0.10)	(0.02)	(0.11)
Republican	0.63***	0.66***	0.59***	0.61***
	(0.02)	(0.09)	(0.03)	(0.10)
Democrat x Pro-Democrat Headline	0.02	0.02	0.01	0.01
	(0.03)	(0.03)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.01	0.01	−0.03	−0.02
	(0.02)	(0.03)	(0.03)	(0.03)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.03		−0.01
		(0.03)		(0.03)
Education		−0.01		0.00
		(0.01)		(0.01)
Hispanic		−0.10*		−0.08·
		(0.04)		(0.04)
Race: Non-White		−0.02		−0.04
		(0.03)		(0.03)
Household Income		0.00		0.00
		(0.01)		(0.01)
Political Interest		−0.00		−0.01
		(0.02)		(0.02)
Social Media Most Common News Format		−0.01		−0.02
		(0.04)		(0.04)
Social Media Post Flagged		−0.01		−0.02
		(0.04)		(0.05)
Social Media Post Removed		0.12**		0.13*
		(0.04)		(0.05)
R^2	0.49	0.50	0.44	0.45
Adj. R^2	0.49	0.49	0.44	0.44
Num. obs.	2240	2240	1782	1782
N Clusters	1120	1120	1035	1035

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A27. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.30***	0.30**	0.27***	0.28*
	(0.02)	(0.10)	(0.02)	(0.12)
Republican	0.63***	0.62***	0.59***	0.58***
	(0.03)	(0.10)	(0.03)	(0.11)
Democrat x Pro-Democrat Headline	0.02	0.02	0.01	0.01
	(0.03)	(0.03)	(0.04)	(0.04)
Republican x Pro-Republican Headline	0.00	0.00	−0.03	−0.03
	(0.04)	(0.04)	(0.05)	(0.05)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		0.01
		(0.03)		(0.03)
Education		0.00		0.01
		(0.01)		(0.01)
Hispanic		−0.11**		−0.08·
		(0.04)		(0.04)
Race: Non-White		0.01		−0.01
		(0.03)		(0.03)
Household Income		−0.00		−0.00
		(0.01)		(0.01)
Political Interest		−0.00		−0.02
		(0.02)		(0.02)
Social Media Most Common News Format		0.00		0.00
		(0.04)		(0.05)
Social Media Post Flagged		−0.01		−0.00
		(0.06)		(0.06)
Social Media Post Removed		0.11·		0.11
		(0.06)		(0.07)
R^2	0.52	0.53	0.47	0.48
Adj. R^2	0.52	0.52	0.47	0.47
Num. obs.	2240	2240	1782	1782

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A28. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.28***	0.35***	0.26***	0.31**
	(0.02)	(0.08)	(0.02)	(0.09)
Republican	0.63***	0.66***	0.59***	0.61***
	(0.02)	(0.08)	(0.03)	(0.09)
Democrat x Pro-Democrat Headline	0.02	0.02	0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.03)
Republican x Pro-Republican Headline	0.01	0.01	−0.03	−0.02
	(0.03)	(0.03)	(0.04)	(0.04)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.03		−0.01
		(0.02)		(0.03)
Education		−0.01		0.00
		(0.01)		(0.01)
Hispanic		−0.10**		−0.08*
		(0.03)		(0.04)
Race: Non-White		−0.02		−0.04·
		(0.02)		(0.03)
Household Income		0.00		0.00
		(0.01)		(0.01)
Political Interest		−0.00		−0.01
		(0.02)		(0.02)
Social Media Most Common News Format		−0.01		−0.02
		(0.03)		(0.04)
Social Media Post Flagged		−0.01		−0.02
		(0.04)		(0.04)
Social Media Post Removed		0.12**		0.13**
		(0.04)		(0.05)
R^2	0.49	0.50	0.44	0.45
Adj. R^2	0.49	0.49	0.44	0.44
Num. obs.	2240	2240	1782	1782

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

A2.1.2. Main Models With All Headlines and Pre-Registered Censorship Coding

Perception of Headline Removal as Censorship

Results Based on Listwise Deletion

Weighted Model Using Clustered SEs

Table A29. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.22***	−0.05	0.20***	−0.08
	(0.02)	(0.13)	(0.02)	(0.13)
Republican	0.53***	0.28*	0.48***	0.23·
	(0.03)	(0.12)	(0.03)	(0.13)
Democrat x Pro-Democrat Headline	0.01	0.03	0.02	0.03
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.00	0.01	−0.03	−0.02
	(0.02)	(0.03)	(0.03)	(0.04)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.00		0.02
		(0.03)		(0.04)
Education		0.01		0.02
		(0.02)		(0.02)
Hispanic		−0.05		−0.02
		(0.05)		(0.05)
Race: Non-White		0.03		0.04
		(0.04)		(0.04)
Household Income		0.00		−0.00
		(0.01)		(0.01)
Political Interest		0.05·		0.05
		(0.03)		(0.03)
Social Media Most Common News Format		−0.01		−0.01
		(0.05)		(0.05)
Social Media Post Flagged		−0.08		−0.10
		(0.06)		(0.07)
Social Media Post Removed		0.18*		0.21*
		(0.07)		(0.09)
R^2	0.43	0.46	0.38	0.40
Adj. R^2	0.43	0.45	0.38	0.40
Num. obs.	2203	1697	1725	1351
RMSE	0.46	0.46	0.44	0.44
N Clusters	1110	853	1007	787

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Results Based on Multiple Imputation

Weighted Model Using Clustered SEs

Table A30. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.22***	0.07	0.20***	0.07
	(0.02)	(0.11)	(0.02)	(0.12)
Republican	0.53***	0.37**	0.49***	0.35**
	(0.03)	(0.11)	(0.03)	(0.12)
Democrat x Pro-Democrat Headline	0.02	0.02	0.02	0.02
	(0.02)	(0.02)	(0.02)	(0.02)
Republican x Pro-Republican Headline	0.00	0.00	−0.03	−0.03
	(0.03)	(0.03)	(0.03)	(0.03)
Age		−0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		0.01
		(0.03)		(0.03)
Education		0.02		0.02
		(0.02)		(0.02)
Hispanic		−0.05		−0.02
		(0.04)		(0.05)
Race: Non-White		−0.01		−0.01
		(0.03)		(0.04)
Household Income		0.00		0.00
		(0.01)		(0.01)
Political Interest		0.03		0.01
		(0.02)		(0.03)
Social Media Most Common News Format		−0.00		−0.01
		(0.05)		(0.05)
Social Media Post Flagged		−0.03		−0.02
		(0.05)		(0.06)
Social Media Post Removed		0.12*		0.12·
		(0.05)		(0.06)
R^2	0.43	0.44	0.38	0.39
Adj. R^2	0.43	0.44	0.38	0.38
Num. obs.	2240	2240	1771	1771
N Clusters	1120	1120	1033	1033

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

A2.1.3. Main Models Considering Only the First Headline

Intent to Remove Headline

Results Based on Listwise Deletion

Weighted Model Using Clustered SEs

Table A31. Regression of Intent to Remove Headline on Partisanship and Alignment

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.78***	0.86***	0.80***	0.99***
	(0.03)	(0.14)	(0.03)	(0.15)
Republican	0.38***	0.43**	0.42***	0.62***
	(0.04)	(0.14)	(0.04)	(0.16)
Democrat x Pro-Democrat Headline	−0.12**	−0.13**	−0.08·	−0.09·
	(0.04)	(0.05)	(0.05)	(0.05)
Republican x Pro-Republican Headline	−0.02	0.01	0.04	0.06
	(0.05)	(0.06)	(0.06)	(0.07)
Age		0.00		−0.00
		(0.00)		(0.00)
Gender: Female		0.00		0.02
		(0.04)		(0.04)
Education		−0.01		−0.04·
		(0.02)		(0.02)
Hispanic		0.07		0.08
		(0.05)		(0.06)
Race: Non-White		−0.02		0.03
		(0.04)		(0.05)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.01		0.01
		(0.03)		(0.04)
Social Media Most Common News Format		0.03		0.03
		(0.06)		(0.06)
Social Media Post Flagged		0.03		0.08
		(0.08)		(0.10)
Social Media Post Removed		−0.15·		−0.25*
		(0.08)		(0.09)
R^2	0.61	0.61	0.67	0.68
Adj. R^2	0.61	0.61	0.66	0.68
Num. obs.	1096	844	864	675
RMSE	0.46	0.46	0.45	0.45
N Clusters	1096	844	864	675

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A32. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.79***	0.78***	0.82***	0.92***
	(0.02)	(0.11)	(0.02)	(0.12)
Republican	0.39***	0.37**	0.43***	0.55***
	(0.03)	(0.11)	(0.04)	(0.13)
Democrat x Pro-Democrat Headline	−0.10**	−0.11**	−0.06·	−0.07·
	(0.03)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	−0.04	−0.02	0.03	0.03
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.03		0.05
		(0.03)		(0.03)
Education		0.00		−0.03
		(0.02)		(0.02)
Hispanic		0.08		0.08
		(0.05)		(0.05)
Race: Non-White		0.01		0.03
		(0.04)		(0.04)
Household Income		−0.01		−0.02
		(0.01)		(0.01)
Political Interest		−0.01		−0.00
		(0.03)		(0.03)
Social Media Most Common News Format		0.03		0.05
		(0.04)		(0.05)
Social Media Post Flagged		0.03		0.06
		(0.06)		(0.07)
Social Media Post Removed		−0.18**		−0.24**
		(0.06)		(0.07)
R^2	0.65	0.65	0.70	0.72
Adj. R^2	0.65	0.65	0.70	0.71
Num. obs.	1096	844	864	675
RMSE	0.46	0.45	0.44	0.44
N Clusters	1096	844	864	675

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A33. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.78***	0.86***	0.80***	0.99***
	(0.03)	(0.14)	(0.03)	(0.15)
Republican	0.38***	0.43**	0.42***	0.62***
	(0.04)	(0.14)	(0.04)	(0.16)
Democrat x Pro-Democrat Headline	−0.12**	−0.13**	−0.08·	−0.09·
	(0.04)	(0.05)	(0.05)	(0.05)
Republican x Pro-Republican Headline	−0.02	0.01	0.04	0.06
	(0.05)	(0.06)	(0.06)	(0.07)
Age		0.00		−0.00
		(0.00)		(0.00)
Gender: Female		0.00		0.02
		(0.04)		(0.04)
Education		−0.01		−0.04·
		(0.02)		(0.02)
Hispanic		0.07		0.08
		(0.05)		(0.06)
Race: Non-White		−0.02		0.03
		(0.04)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		−0.01		0.01
		(0.03)		(0.04)
Social Media Most Common News Format		0.03		0.03
		(0.05)		(0.06)
Social Media Post Flagged		0.03		0.08
		(0.08)		(0.10)
Social Media Post Removed		−0.15·		−0.25**
		(0.08)		(0.09)
R^2	0.61	0.61	0.67	0.68
Adj. R^2	0.61	0.61	0.66	0.68
Num. obs.	1096	844	864	675
RMSE	0.46	0.46	0.45	0.45

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A34. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.79***	0.78***	0.82***	0.92***
	(0.02)	(0.11)	(0.02)	(0.12)
Republican	0.39***	0.37**	0.43***	0.55***
	(0.03)	(0.11)	(0.04)	(0.13)
Democrat x Pro-Democrat Headline	−0.10**	−0.11**	−0.06·	−0.07·
	(0.03)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	−0.04	−0.02	0.03	0.03
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.03		0.05
		(0.03)		(0.03)
Education		0.00		−0.03
		(0.02)		(0.02)
Hispanic		0.08		0.08
		(0.05)		(0.05)
Race: Non-White		0.01		0.03
		(0.04)		(0.04)
Household Income		−0.01		−0.02
		(0.01)		(0.01)
Political Interest		−0.01		−0.00
		(0.03)		(0.03)
Social Media Most Common News Format		0.03		0.05
		(0.04)		(0.05)
Social Media Post Flagged		0.03		0.06
		(0.06)		(0.07)
Social Media Post Removed		−0.18**		−0.24**
		(0.06)		(0.07)
R^2	0.65	0.65	0.70	0.72
Adj. R^2	0.65	0.65	0.70	0.71
Num. obs.	1096	844	864	675
RMSE	0.46	0.45	0.44	0.44

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Results Based on Multiple Imputation

Weighted Model Using Clustered SEs

TABLE A35. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.78***	0.74***	0.80***	0.80***
	(0.03)	(0.12)	(0.03)	(0.14)
Republican	0.37***	0.35**	0.42***	0.48**
	(0.04)	(0.13)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.11**	−0.11**	−0.09·	−0.09·
	(0.04)	(0.04)	(0.05)	(0.04)
Republican x Pro-Republican Headline	−0.01	−0.02	0.04	0.02
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.02		0.04
		(0.03)		(0.04)
Education		−0.01		−0.03
		(0.02)		(0.02)
Hispanic		0.09·		0.12*
		(0.05)		(0.05)
Race: Non-White		−0.01		0.04
		(0.04)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		0.00		0.04
		(0.03)		(0.03)
Social Media Most Common News Format		0.01		0.00
		(0.05)		(0.06)
Social Media Post Flagged		0.00		0.01
		(0.06)		(0.08)
Social Media Post Removed		−0.11		−0.15
		(0.07)		(0.09)
R^2	0.61	0.62	0.66	0.67
Adj. R^2	0.61	0.61	0.66	0.67
Num. obs.	1120	1120	893	893
N Clusters	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A36. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.79***	0.68***	0.81***	0.78***
	(0.02)	(0.10)	(0.02)	(0.11)
Republican	0.39***	0.30**	0.43***	0.45***
	(0.03)	(0.10)	(0.04)	(0.11)
Democrat x Pro-Democrat Headline	−0.11**	−0.11**	−0.07·	−0.07*
	(0.03)	(0.03)	(0.04)	(0.04)
Republican x Pro-Republican Headline	−0.03	−0.04	0.03	0.01
	(0.05)	(0.05)	(0.06)	(0.05)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.05·		0.07*
		(0.03)		(0.03)
Education		0.00		−0.02
		(0.01)		(0.02)
Hispanic		0.09*		0.12*
		(0.04)		(0.05)
Race: Non-White		0.02		0.06
		(0.03)		(0.03)
Household Income		−0.00		−0.02
		(0.01)		(0.01)
Political Interest		0.00		0.02
		(0.02)		(0.03)
Social Media Most Common News Format		0.01		0.02
		(0.04)		(0.05)
Social Media Post Flagged		0.01		0.01
		(0.05)		(0.06)
Social Media Post Removed		−0.14*		−0.16·
		(0.06)		(0.08)
R^2	0.65	0.65	0.70	0.71
Adj. R^2	0.64	0.65	0.70	0.70
Num. obs.	1120	1120	893	893
N Clusters	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A37. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.78***	0.74***	0.80***	0.80***
	(0.03)	(0.12)	(0.03)	(0.14)
Republican	0.37***	0.35**	0.42***	0.48**
	(0.04)	(0.13)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.11**	−0.11**	−0.09·	−0.09·
	(0.04)	(0.04)	(0.05)	(0.04)
Republican x Pro-Republican Headline	−0.01	−0.02	0.04	0.02
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.02		0.04
		(0.03)		(0.04)
Education		−0.01		−0.03
		(0.02)		(0.02)
Hispanic		0.09*		0.12*
		(0.05)		(0.05)
Race: Non-White		−0.01		0.04
		(0.04)		(0.04)
Household Income		−0.00		−0.01
		(0.01)		(0.01)
Political Interest		0.00		0.04
		(0.03)		(0.03)
Social Media Most Common News Format		0.01		0.00
		(0.05)		(0.06)
Social Media Post Flagged		0.00		0.01
		(0.06)		(0.08)
Social Media Post Removed		−0.11·		−0.15
		(0.07)		(0.09)
R^2	0.61	0.62	0.66	0.67
Adj. R^2	0.61	0.61	0.66	0.67
Num. obs.	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A38. REGRESSION OF INTENT TO REMOVE HEADLINE ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.79***	0.68***	0.81***	0.78***
	(0.02)	(0.10)	(0.02)	(0.11)
Republican	0.39***	0.30**	0.43***	0.45***
	(0.03)	(0.10)	(0.04)	(0.11)
Democrat x Pro-Democrat Headline	−0.11**	−0.11**	−0.07·	−0.07*
	(0.03)	(0.03)	(0.04)	(0.04)
Republican x Pro-Republican Headline	−0.03	−0.04	0.03	0.01
	(0.05)	(0.05)	(0.06)	(0.05)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.05·		0.07*
		(0.03)		(0.03)
Education		0.00		−0.02
		(0.01)		(0.02)
Hispanic		0.09*		0.12*
		(0.04)		(0.05)
Race: Non-White		0.02		0.06
		(0.03)		(0.03)
Household Income		−0.00		−0.02
		(0.01)		(0.01)
Political Interest		0.00		0.02
		(0.02)		(0.03)
Social Media Most Common News Format		0.01		0.02
		(0.04)		(0.05)
Social Media Post Flagged		0.01		0.01
		(0.05)		(0.06)
Social Media Post Removed		−0.14*		−0.16·
		(0.06)		(0.08)
R^2	0.65	0.65	0.70	0.71
Adj. R^2	0.64	0.65	0.70	0.70
Num. obs.	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Intent to Report Headline as Harmful

Results Based on Listwise Deletion

Weighted Model Using Clustered SEs

Table A39. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.63***	0.73***	0.64***	0.76***
	(0.03)	(0.13)	(0.03)	(0.15)
Republican	0.27***	0.35*	0.31***	0.40*
	(0.03)	(0.13)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.18***	−0.18***	−0.15**	−0.15*
	(0.05)	(0.05)	(0.05)	(0.06)
Republican x Pro-Republican Headline	0.05	0.06	0.08	0.08
	(0.05)	(0.05)	(0.06)	(0.07)
Age		0.00·		0.00
		(0.00)		(0.00)
Gender: Female		−0.03		−0.02
		(0.04)		(0.04)
Education		−0.01		−0.03
		(0.02)		(0.02)
Hispanic		−0.03		−0.02
		(0.06)		(0.07)
Race: Non-White		0.00		−0.02
		(0.04)		(0.05)
Household Income		−0.01		−0.02
		(0.01)		(0.02)
Political Interest		−0.04		0.00
		(0.03)		(0.04)
Social Media Most Common News Format		0.05		0.06
		(0.05)		(0.06)
Social Media Post Flagged		−0.04		−0.03
		(0.06)		(0.09)
Social Media Post Removed		−0.07		−0.06
		(0.07)		(0.09)
R^2	0.47	0.47	0.51	0.51
Adj. R^2	0.47	0.46	0.51	0.50
Num. obs.	1094	845	862	674
RMSE	0.47	0.47	0.47	0.47
N Clusters	1094	845	862	674

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A40. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.62***	0.62***	0.64***	0.63***
	(0.03)	(0.12)	(0.03)	(0.13)
Republican	0.27***	0.27*	0.31***	0.30*
	(0.03)	(0.11)	(0.04)	(0.13)
Democrat x Pro-Democrat Headline	−0.16***	−0.15**	−0.14***	−0.13**
	(0.04)	(0.04)	(0.04)	(0.05)
Republican x Pro-Republican Headline	0.05	0.04	0.08	0.06
	(0.04)	(0.05)	(0.05)	(0.06)
Age		0.00*		0.00·
		(0.00)		(0.00)
Gender: Female		−0.03		−0.02
		(0.03)		(0.04)
Education		−0.01		−0.03
		(0.02)		(0.02)
Hispanic		−0.03		−0.01
		(0.05)		(0.06)
Race: Non-White		0.02		0.02
		(0.04)		(0.04)
Household Income		−0.01		−0.01
		(0.01)		(0.01)
Political Interest		−0.02		0.01
		(0.03)		(0.03)
Social Media Most Common News Format		0.00		0.01
		(0.05)		(0.05)
Social Media Post Flagged		−0.01		−0.01
		(0.06)		(0.07)
Social Media Post Removed		−0.08		−0.07
		(0.06)		(0.08)
R^2	0.49	0.48	0.53	0.52
Adj. R^2	0.49	0.47	0.52	0.51
Num. obs.	1094	845	862	674
RMSE	0.48	0.48	0.48	0.48
N Clusters	1094	845	862	674

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A41. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.63***	0.73***	0.64***	0.76***
	(0.03)	(0.13)	(0.03)	(0.15)
Republican	0.27***	0.35**	0.31***	0.40**
	(0.03)	(0.13)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.18***	−0.18***	−0.15**	−0.15*
	(0.05)	(0.05)	(0.05)	(0.06)
Republican x Pro-Republican Headline	0.05	0.06	0.08	0.08
	(0.05)	(0.05)	(0.06)	(0.07)
Age		0.00*		0.00
		(0.00)		(0.00)
Gender: Female		−0.03		−0.02
		(0.04)		(0.04)
Education		−0.01		−0.03
		(0.02)		(0.02)
Hispanic		−0.03		−0.02
		(0.06)		(0.07)
Race: Non-White		0.00		−0.02
		(0.04)		(0.05)
Household Income		−0.01		−0.02
		(0.01)		(0.02)
Political Interest		−0.04		0.00
		(0.03)		(0.04)
Social Media Most Common News Format		0.05		0.06
		(0.05)		(0.06)
Social Media Post Flagged		−0.04		−0.03
		(0.06)		(0.09)
Social Media Post Removed		−0.07		−0.06
		(0.06)		(0.09)
R^2	0.47	0.47	0.51	0.51
Adj. R^2	0.47	0.46	0.51	0.50
Num. obs.	1094	845	862	674
RMSE	0.47	0.47	0.47	0.47

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

Table A42. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.62***	0.62***	0.64***	0.63***
	(0.03)	(0.12)	(0.03)	(0.13)
Republican	0.27***	0.27*	0.31***	0.30*
	(0.03)	(0.11)	(0.04)	(0.13)
Democrat x Pro-Democrat Headline	−0.16***	−0.15**	−0.14***	−0.13**
	(0.04)	(0.04)	(0.04)	(0.05)
Republican x Pro-Republican Headline	0.05	0.04	0.08	0.06
	(0.04)	(0.05)	(0.05)	(0.06)
Age		0.00*		0.00·
		(0.00)		(0.00)
Gender: Female		−0.03		−0.02
		(0.03)		(0.04)
Education		−0.01		−0.03
		(0.02)		(0.02)
Hispanic		−0.03		−0.01
		(0.05)		(0.06)
Race: Non-White		0.02		0.02
		(0.04)		(0.04)
Household Income		−0.01		−0.01
		(0.01)		(0.01)
Political Interest		−0.02		0.01
		(0.03)		(0.03)
Social Media Most Common News Format		0.00		0.01
		(0.05)		(0.05)
Social Media Post Flagged		−0.01		−0.01
		(0.06)		(0.07)
Social Media Post Removed		−0.08		−0.07
		(0.06)		(0.08)
R^2	0.49	0.48	0.53	0.52
Adj. R^2	0.49	0.47	0.52	0.51
Num. obs.	1094	845	862	674
RMSE	0.48	0.48	0.48	0.48

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Results Based on Multiple Imputation

Weighted Model Using Clustered SEs

Table A43. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.63***	0.63***	0.64***	0.62***
	(0.03)	(0.12)	(0.03)	(0.14)
Republican	0.27***	0.27*	0.31***	0.30*
	(0.03)	(0.12)	(0.04)	(0.14)
Democrat x Pro-Democrat Headline	−0.19***	−0.19***	−0.16**	−0.17**
	(0.05)	(0.05)	(0.05)	(0.05)
Republican x Pro-Republican Headline	0.05	0.04	0.07	0.06
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00·		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		−0.01
		(0.03)		(0.04)
Education		−0.01		−0.02
		(0.02)		(0.02)
Hispanic		−0.00		0.01
		(0.05)		(0.06)
Race: Non-White		−0.02		−0.03
		(0.04)		(0.04)
Household Income		−0.01		−0.02
		(0.01)		(0.01)
Political Interest		0.00		0.05
		(0.03)		(0.03)
Social Media Most Common News Format		0.04		0.06
		(0.05)		(0.06)
Social Media Post Flagged		−0.05		−0.04
		(0.06)		(0.07)
Social Media Post Removed		−0.06		−0.06
		(0.07)		(0.08)
R^2	0.47	0.48	0.51	0.52
Adj. R^2	0.47	0.47	0.50	0.51
Num. obs.	1120	1120	893	893
N Clusters	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

Table A44. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.63***	0.54***	0.64***	0.55***
	(0.03)	(0.10)	(0.03)	(0.12)
Republican	0.28***	0.19·	0.31***	0.22·
	(0.03)	(0.10)	(0.04)	(0.12)
Democrat x Pro-Democrat Headline	−0.16***	−0.16***	−0.15***	−0.15***
	(0.04)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	0.05	0.04	0.07	0.06
	(0.04)	(0.04)	(0.05)	(0.05)
Age		0.00*		0.00·
		(0.00)		(0.00)
Gender: Female		−0.01		−0.01
		(0.03)		(0.03)
Education		−0.01		−0.02
		(0.01)		(0.02)
Hispanic		−0.01		0.01
		(0.04)		(0.05)
Race: Non-White		0.01		0.01
		(0.03)		(0.04)
Household Income		−0.01		−0.02
		(0.01)		(0.01)
Political Interest		0.02		0.05·
		(0.02)		(0.03)
Social Media Most Common News Format		−0.01		0.00
		(0.04)		(0.05)
Social Media Post Flagged		−0.03		−0.03
		(0.05)		(0.06)
Social Media Post Removed		−0.05		−0.05
		(0.06)		(0.07)
R^2	0.49	0.50	0.52	0.53
Adj. R^2	0.49	0.49	0.52	0.53
Num. obs.	1120	1120	893	893
N Clusters	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A45. REGRESSION OF INTENT TO REPORT HEADLINE AS HARMFUL ON PARTISANSHIP AND ALIGNMENT

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.63***	0.63***	0.64***	0.62***
	(0.03)	(0.12)	(0.03)	(0.14)
Republican	0.27***	0.27*	0.31***	0.30*
	(0.03)	(0.12)	(0.04)	(0.13)
Democrat x Pro-Democrat Headline	−0.19***	−0.19***	−0.16**	−0.17***
	(0.05)	(0.05)	(0.05)	(0.05)
Republican x Pro-Republican Headline	0.05	0.04	0.07	0.06
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00*		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		−0.01
		(0.03)		(0.04)
Education		−0.01		−0.02
		(0.02)		(0.02)
Hispanic		−0.00		0.01
		(0.05)		(0.06)
Race: Non-White		−0.02		−0.03
		(0.04)		(0.04)
Household Income		−0.01		−0.02
		(0.01)		(0.01)
Political Interest		0.00		0.05
		(0.03)		(0.03)
Social Media Most Common News Format		0.04		0.06
		(0.05)		(0.06)
Social Media Post Flagged		−0.05		−0.04
		(0.06)		(0.07)
Social Media Post Removed		−0.06		−0.06
		(0.07)		(0.08)
R^2	0.47	0.48	0.51	0.52
Adj. R^2	0.47	0.47	0.50	0.51
Num. obs.	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

Table A46. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.63***	0.54***	0.64***	0.55***
	(0.03)	(0.10)	(0.03)	(0.12)
Republican	0.28***	0.19·	0.31***	0.22·
	(0.03)	(0.10)	(0.04)	(0.12)
Democrat x Pro-Democrat Headline	−0.16***	−0.16***	−0.15***	−0.15***
	(0.04)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	0.05	0.04	0.07	0.06
	(0.04)	(0.04)	(0.05)	(0.05)
Age		0.00*		0.00·
		(0.00)		(0.00)
Gender: Female		−0.01		−0.01
		(0.03)		(0.03)
Education		−0.01		−0.02
		(0.01)		(0.02)
Hispanic		−0.01		0.01
		(0.04)		(0.05)
Race: Non-White		0.01		0.01
		(0.03)		(0.04)
Household Income		−0.01		−0.02
		(0.01)		(0.01)
Political Interest		0.02		0.05·
		(0.02)		(0.03)
Social Media Most Common News Format		−0.01		0.00
		(0.04)		(0.05)
Social Media Post Flagged		−0.03		−0.03
		(0.05)		(0.06)
Social Media Post Removed		−0.05		−0.05
		(0.06)		(0.07)
R^2	0.49	0.50	0.52	0.53
Adj. R^2	0.49	0.49	0.52	0.53
Num. obs.	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Perception of Headline Removal as Censorship

Results Based on Listwise Deletion

Weighted Model Using Clustered SEs

Table A47. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.30***	0.07	0.28***	0.02
	(0.03)	(0.17)	(0.04)	(0.20)
Republican	0.62***	0.40*	0.56***	0.31
	(0.04)	(0.16)	(0.05)	(0.19)
Democrat x Pro-Democrat Headline	−0.02	−0.01	−0.03	−0.02
	(0.05)	(0.06)	(0.05)	(0.06)
Republican x Pro-Republican Headline	0.08	0.10	0.07	0.08
	(0.06)	(0.06)	(0.07)	(0.08)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.01		0.02
		(0.04)		(0.05)
Education		0.01		0.02
		(0.02)		(0.03)
Hispanic		−0.06		−0.02
		(0.06)		(0.07)
Race: Non-White		0.01		0.03
		(0.05)		(0.05)
Household Income		−0.00		0.00
		(0.01)		(0.02)
Political Interest		0.04		0.02
		(0.04)		(0.04)
Social Media Most Common News Format		0.02		−0.00
		(0.06)		(0.07)
Social Media Post Flagged		−0.07		−0.16·
		(0.07)		(0.08)
Social Media Post Removed		0.17*		0.29**
		(0.07)		(0.09)
R^2	0.54	0.56	0.47	0.50
Adj. R^2	0.54	0.55	0.47	0.49
Num. obs.	887	705	700	561
RMSE	0.46	0.46	0.46	0.45
N Clusters	887	705	700	561

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A48. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.29***	0.15	0.26***	0.11
	(0.03)	(0.13)	(0.03)	(0.15)
Republican	0.62***	0.48***	0.57***	0.41**
	(0.04)	(0.13)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.01	−0.01	−0.01	−0.02
	(0.04)	(0.04)	(0.04)	(0.05)
Republican x Pro-Republican Headline	0.07	0.10·	0.05	0.07
	(0.05)	(0.05)	(0.06)	(0.07)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.00		−0.01
		(0.04)		(0.04)
Education		−0.01		−0.00
		(0.02)		(0.02)
Hispanic		−0.05		−0.04
		(0.06)		(0.06)
Race: Non-White		0.01		0.02
		(0.04)		(0.05)
Household Income		0.01		0.01
		(0.01)		(0.01)
Political Interest		0.03		0.02
		(0.03)		(0.03)
Social Media Most Common News Format		−0.05		−0.07
		(0.05)		(0.05)
Social Media Post Flagged		−0.05		−0.09
		(0.06)		(0.08)
Social Media Post Removed		0.15*		0.23**
		(0.06)		(0.08)
R^2	0.52	0.55	0.45	0.49
Adj. R^2	0.52	0.54	0.45	0.48
Num. obs.	887	705	700	561
RMSE	0.46	0.45	0.46	0.45
N Clusters	887	705	700	561

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

Table A49. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.30***	0.07	0.28***	0.02
	(0.03)	(0.17)	(0.04)	(0.19)
Republican	0.62***	0.40*	0.56***	0.31·
	(0.04)	(0.16)	(0.05)	(0.19)
Democrat x Pro-Democrat Headline	−0.02	−0.01	−0.03	−0.02
	(0.05)	(0.06)	(0.05)	(0.06)
Republican x Pro-Republican Headline	0.08	0.10	0.07	0.08
	(0.06)	(0.06)	(0.07)	(0.08)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.01		0.02
		(0.04)		(0.05)
Education		0.01		0.02
		(0.02)		(0.03)
Hispanic		−0.06		−0.02
		(0.06)		(0.07)
Race: Non-White		0.01		0.03
		(0.05)		(0.05)
Household Income		−0.00		0.00
		(0.01)		(0.02)
Political Interest		0.04		0.02
		(0.04)		(0.04)
Social Media Most Common News Format		0.02		−0.00
		(0.06)		(0.07)
Social Media Post Flagged		−0.07		−0.16*
		(0.07)		(0.08)
Social Media Post Removed		0.17*		0.29***
		(0.07)		(0.09)
R^2	0.54	0.56	0.47	0.50
Adj. R^2	0.54	0.55	0.47	0.49
Num. obs.	887	705	700	561
RMSE	0.46	0.46	0.46	0.45

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A50. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.29***	0.15	0.26***	0.11
	(0.03)	(0.13)	(0.03)	(0.15)
Republican	0.62***	0.48***	0.57***	0.41**
	(0.04)	(0.13)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.01	−0.01	−0.01	−0.02
	(0.04)	(0.04)	(0.04)	(0.05)
Republican x Pro-Republican Headline	0.07	0.10.	0.05	0.07
	(0.05)	(0.05)	(0.06)	(0.07)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.00		−0.01
		(0.04)		(0.04)
Education		−0.01		−0.00
		(0.02)		(0.02)
Hispanic		−0.05		−0.04
		(0.06)		(0.06)
Race: Non-White		0.01		0.02
		(0.04)		(0.05)
Household Income		0.01		0.01
		(0.01)		(0.01)
Political Interest		0.03		0.02
		(0.03)		(0.03)
Social Media Most Common News Format		−0.05		−0.07
		(0.05)		(0.05)
Social Media Post Flagged		−0.05		−0.09
		(0.06)		(0.08)
Social Media Post Removed		0.15*		0.23**
		(0.06)		(0.08)
R^2	0.52	0.55	0.45	0.49
Adj. R^2	0.52	0.54	0.45	0.48
Num. obs.	887	705	700	561
RMSE	0.46	0.45	0.46	0.45

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{.}p < 0.1$

Results Based on Multiple Imputation

Weighted Model Using Clustered SEs

Table A51. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.30***	0.24·	0.28***	0.20
	(0.03)	(0.14)	(0.03)	(0.17)
Republican	0.60***	0.53**	0.56***	0.45*
	(0.04)	(0.15)	(0.05)	(0.17)
Democrat x Pro-Democrat Headline	0.01	0.01	−0.02	−0.02
	(0.05)	(0.05)	(0.05)	(0.05)
Republican x Pro-Republican Headline	0.07	0.07	0.04	0.04
	(0.05)	(0.05)	(0.07)	(0.07)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.00		0.00
		(0.04)		(0.04)
Education		0.02		0.02
		(0.02)		(0.02)
Hispanic		−0.08		−0.05
		(0.05)		(0.06)
Race: Non-White		0.00		−0.01
		(0.04)		(0.05)
Household Income		−0.00		0.00
		(0.01)		(0.01)
Political Interest		0.00		−0.02
		(0.03)		(0.03)
Social Media Most Common News Format		0.01		0.02
		(0.06)		(0.06)
Social Media Post Flagged		−0.03		−0.05
		(0.07)		(0.09)
Social Media Post Removed		0.14·		0.15
		(0.08)		(0.10)
R^2	0.53	0.53	0.46	0.47
Adj. R^2	0.52	0.53	0.46	0.46
Num. obs.	1120	1120	893	893
N Clusters	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Clustered SEs

TABLE A52. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.28***	0.31**	0.27***	0.26*
	(0.02)	(0.11)	(0.03)	(0.12)
Republican	0.60***	0.60***	0.56***	0.52***
	(0.04)	(0.11)	(0.04)	(0.13)
Democrat x Pro-Democrat Headline	0.02	0.02	−0.00	−0.01
	(0.04)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	0.07	0.07	0.03	0.04
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		−0.01
		(0.03)		(0.03)
Education		−0.00		0.01
		(0.02)		(0.02)
Hispanic		−0.07		−0.06
		(0.05)		(0.05)
Race: Non-White		−0.02		−0.04
		(0.03)		(0.04)
Household Income		0.00		0.00
		(0.01)		(0.01)
Political Interest		−0.00		−0.01
		(0.02)		(0.03)
Social Media Most Common News Format		−0.02		−0.02
		(0.04)		(0.05)
Social Media Post Flagged		−0.02		−0.04
		(0.06)		(0.07)
Social Media Post Removed		0.14*		0.16*
		(0.06)		(0.07)
R^2	0.50	0.51	0.44	0.45
Adj. R^2	0.50	0.50	0.44	0.44
Num. obs.	1120	1120	893	893
N Clusters	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Weighted Model Using Non-Clustered SEs

TABLE A53. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.30***	0.24·	0.28***	0.20
	(0.03)	(0.14)	(0.03)	(0.16)
Republican	0.60***	0.53**	0.56***	0.45*
	(0.04)	(0.15)	(0.05)	(0.17)
Democrat x Pro-Democrat Headline	0.01	0.01	−0.02	−0.02
	(0.05)	(0.05)	(0.05)	(0.05)
Republican x Pro-Republican Headline	0.07	0.07	0.04	0.04
	(0.05)	(0.05)	(0.07)	(0.07)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.00		0.00
		(0.04)		(0.04)
Education		0.02		0.02
		(0.02)		(0.02)
Hispanic		−0.08		−0.05
		(0.05)		(0.06)
Race: Non-White		0.00		−0.01
		(0.04)		(0.05)
Household Income		−0.00		0.00
		(0.01)		(0.01)
Political Interest		0.00		−0.02
		(0.03)		(0.03)
Social Media Most Common News Format		0.01		0.02
		(0.06)		(0.06)
Social Media Post Flagged		−0.03		−0.05
		(0.07)		(0.09)
Social Media Post Removed		0.14·		0.15
		(0.07)		(0.10)
R^2	0.53	0.53	0.46	0.47
Adj. R^2	0.52	0.53	0.46	0.46
Num. obs.	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Unweighted Model Using Non-Clustered SEs

TABLE A54. REGRESSION OF PERCEPTION OF HEADLINE REMOVAL AS CENSORSHIP ON PARTISANSHIP AND ALIGNMENT

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.28***	0.31**	0.27***	0.26*
	(0.02)	(0.11)	(0.03)	(0.12)
Republican	0.60***	0.60***	0.56***	0.52***
	(0.04)	(0.11)	(0.04)	(0.13)
Democrat x Pro-Democrat Headline	0.02	0.02	−0.00	−0.01
	(0.04)	(0.04)	(0.04)	(0.04)
Republican x Pro-Republican Headline	0.07	0.07	0.03	0.04
	(0.05)	(0.05)	(0.06)	(0.06)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		−0.01		−0.01
		(0.03)		(0.03)
Education		−0.00		0.01
		(0.02)		(0.02)
Hispanic		−0.07		−0.06
		(0.05)		(0.05)
Race: Non-White		−0.02		−0.04
		(0.03)		(0.04)
Household Income		0.00		0.00
		(0.01)		(0.01)
Political Interest		−0.00		−0.01
		(0.02)		(0.03)
Social Media Most Common News Format		−0.02		−0.02
		(0.04)		(0.05)
Social Media Post Flagged		−0.02		−0.04
		(0.06)		(0.07)
Social Media Post Removed		0.14*		0.16*
		(0.06)		(0.07)
R^2	0.50	0.51	0.44	0.45
Adj. R^2	0.50	0.50	0.44	0.44
Num. obs.	1120	1120	893	893

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

A2.1.4. Models Without Interaction

Intent to Remove Headline

Table A55. Regression of Intent to Remove Headline on Partisanship and Alignment

	DV: Intent to Remove Headline	
	All	Inaccurate Subgroup
Democrat	0.69***	0.76***
	(0.02)	(0.02)
Republican	0.34***	0.42***
	(0.02)	(0.03)
R^2	0.58	0.65
Adj. R^2	0.58	0.65
Num. obs.	2190	1721
RMSE	0.46	0.45
N Clusters	1104	1003

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Intent to Report Headline as Harmful

Table A56. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful	
	All	Inaccurate Subgroup
Democrat	0.49***	0.52***
	(0.02)	(0.02)
Republican	0.27***	0.31***
	(0.02)	(0.02)
R^2	0.42	0.45
Adj. R^2	0.42	0.45
Num. obs.	2192	1720
RMSE	0.47	0.47
N Clusters	1105	1005

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Perception of Headline Removal as Censorship

Table A57. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship	
	All	Inaccurate Subgroup
Democrat	0.29***	0.26***
	(0.02)	(0.02)
Republican	0.65***	0.59***
	(0.03)	(0.03)
R^2	0.53	0.47
Adj. R^2	0.53	0.47
Num. obs.	1774	1407
RMSE	0.46	0.45
N Clusters	958	851

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

A2.1.5. Models With Consensus Headlines

The following models include only 8 out of 18 headlines that both Republicans and Democrats think are inaccurate (average rating below 2 on a 4-point scale ranging from "Not at all accurate" to "Very accurate"), and where there is little difference in accuracy perception between Democrats and Republicans (mean absolute difference between average Democrat and Republican accuracy rating below 0.5). These headlines are: Pro-Democrat 1, 4, 5, 7; Pro-Republican 3, 5, 7, 9 (see section A1.4 for all headlines).

Intent to Remove Headline

Table A58. Regression of Intent to Remove Headline on Partisanship and Alignment

	DV: Intent to Remove Headline			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.79***	0.96***	0.80***	1.12***
	(0.03)	(0.14)	(0.03)	(0.15)
Republican	0.40***	0.58***	0.42***	0.74***
	(0.04)	(0.14)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.08·	−0.09·	−0.05	−0.05
	(0.04)	(0.05)	(0.04)	(0.05)
Republican x Pro-Republican Headline	−0.02	−0.04	0.02	0.00
	(0.05)	(0.06)	(0.06)	(0.06)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.05		0.03
		(0.05)		(0.05)
Education		−0.06*		−0.08**
		(0.03)		(0.03)
Hispanic		0.12·		0.05
		(0.06)		(0.06)
Race: Non-White		−0.03		0.01
		(0.05)		(0.05)
Household Income		0.00		0.00
		(0.02)		(0.02)
Political Interest		−0.02		−0.03
		(0.04)		(0.04)
Social Media Most Common News Format		−0.01		−0.01
		(0.06)		(0.06)
Social Media Post Flagged		−0.02		0.08
		(0.09)		(0.09)
Social Media Post Removed		−0.18·		−0.26*
		(0.10)		(0.10)
R^2	0.63	0.65	0.66	0.69
Adj. R^2	0.63	0.64	0.66	0.68
Num. obs.	956	744	826	655
RMSE	0.45	0.45	0.44	0.44
N Clusters	742	577	662	524

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Intent to Report Headline as Harmful

Table A59. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.58***	0.52**	0.58***	0.54**
	(0.04)	(0.15)	(0.04)	(0.16)
Republican	0.28***	0.25·	0.29***	0.27·
	(0.04)	(0.14)	(0.04)	(0.15)
Democrat x Pro-Democrat Headline	−0.12**	−0.10·	−0.09·	−0.07
	(0.05)	(0.05)	(0.05)	(0.05)
Republican x Pro-Republican Headline	−0.03	−0.03	−0.02	−0.00
	(0.05)	(0.05)	(0.05)	(0.06)
Age		0.00·		0.00
		(0.00)		(0.00)
Gender: Female		−0.07		−0.08
		(0.05)		(0.05)
Education		−0.02		−0.03
		(0.02)		(0.02)
Hispanic		−0.01		0.00
		(0.07)		(0.08)
Race: Non-White		0.01		−0.02
		(0.05)		(0.06)
Household Income		0.00		−0.00
		(0.02)		(0.02)
Political Interest		−0.01		0.01
		(0.03)		(0.04)
Social Media Most Common News Format		0.06		0.06
		(0.06)		(0.07)
Social Media Post Flagged		−0.02		0.01
		(0.07)		(0.08)
Social Media Post Removed		−0.11		−0.14
		(0.07)		(0.08)
R^2	0.44	0.42	0.45	0.43
Adj. R^2	0.44	0.41	0.45	0.42
Num. obs.	961	747	830	657
RMSE	0.46	0.46	0.46	0.46
N Clusters	748	580	666	526

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Perception of Headline Removal as Censorship

Table A60. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship			
	All		Inaccurate Subgroup	
	Baseline	Controls	Baseline	Controls
Democrat	0.29***	−0.08	0.27***	−0.17
	(0.04)	(0.23)	(0.04)	(0.23)
Republican	0.62***	0.25	0.61***	0.17
	(0.05)	(0.21)	(0.05)	(0.21)
Democrat x Pro-Democrat Headline	−0.06	−0.03	−0.08·	−0.07
	(0.05)	(0.05)	(0.05)	(0.06)
Republican x Pro-Republican Headline	−0.00	0.02	−0.05	−0.03
	(0.05)	(0.06)	(0.06)	(0.07)
Age		0.00		0.00
		(0.00)		(0.00)
Gender: Female		0.01		0.03
		(0.05)		(0.05)
Education		0.01		0.02
		(0.03)		(0.03)
Hispanic		−0.04		−0.01
		(0.08)		(0.08)
Race: Non-White		0.06		0.05
		(0.06)		(0.06)
Household Income		0.02		0.02
		(0.02)		(0.02)
Political Interest		0.03		0.05
		(0.04)		(0.05)
Social Media Most Common News Format		0.03		−0.00
		(0.07)		(0.07)
Social Media Post Flagged		−0.08		−0.12
		(0.09)		(0.09)
Social Media Post Removed		0.23*		0.25*
		(0.11)		(0.11)
R^2	0.51	0.53	0.48	0.50
Adj. R^2	0.51	0.52	0.48	0.49
Num. obs.	783	621	687	552
RMSE	0.45	0.45	0.44	0.44
N Clusters	621	491	559	449

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

A2.1.6. Models Disaggregated by Headline

Intent to Remove Headline

Table A61. Regression of Intent to Remove Headline on Partisanship by Headline (Pro-Democrat Headlines)

	DV: Intent to Remove Headline								
Headline	1	2	3	4	5	6	7	8	9
Democrat Respondent	0.64***	0.40***	0.55***	0.74***	0.70***	0.70***	0.77***	0.64***	0.59***
	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)	(0.08)
Republican Respondent	0.42***	0.49***	0.14**	0.29**	0.34***	0.20**	0.58***	0.51***	0.19**
	(0.08)	(0.08)	(0.05)	(0.08)	(0.08)	(0.06)	(0.10)	(0.08)	(0.05)
R^2	0.54	0.45	0.45	0.65	0.59	0.60	0.70	0.59	0.48
Adj. R^2	0.54	0.44	0.44	0.64	0.58	0.59	0.70	0.58	0.47
Num. obs.	123	128	127	129	124	120	103	122	119
RMSE	0.50	0.49	0.43	0.43	0.45	0.44	0.44	0.50	0.45
N Clusters	123	128	127	129	124	120	103	122	119

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Table A62. Regression of Intent to Remove Headline on Partisanship by Headline (Pro-Republican Headlines)

	DV: Intent to Remove Headline								
Headline	1	2	3	4	5	6	7	8	9
Democrat Respondent	0.67***	0.72***	0.68***	0.81***	0.87***	0.74***	0.91***	0.61***	0.71***
	(0.06)	(0.07)	(0.07)	(0.05)	(0.05)	(0.07)	(0.04)	(0.07)	(0.06)
Republican Respondent	0.23**	0.30***	0.47***	0.40***	0.36***	0.34***	0.42***	0.31***	0.31***
	(0.06)	(0.08)	(0.08)	(0.09)	(0.08)	(0.07)	(0.08)	(0.07)	(0.08)
R^2	0.56	0.63	0.60	0.71	0.72	0.61	0.77	0.50	0.58
Adj. R^2	0.55	0.62	0.60	0.70	0.71	0.60	0.76	0.49	0.57
Num. obs.	137	120	119	112	112	129	115	120	131
RMSE	0.45	0.46	0.47	0.44	0.41	0.47	0.39	0.47	0.46
N Clusters	137	120	119	112	112	129	115	120	131

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Intent to Report Headline as Harmful

Table A63. Regression of Intent to Report Headline as Harmful on Partisanship by Headline (Pro-Democrat Headlines)

	DV: Intent to Report Headline as Harmful								
Headline	1	2	3	4	5	6	7	8	9
Democrat Respondent	0.33***	0.29***	0.54***	0.59***	0.38***	0.48***	0.51***	0.36***	0.37***
	(0.07)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Republican Respondent	0.20**	0.35***	0.13*	0.26**	0.34***	0.25**	0.36**	0.31***	0.18**
	(0.06)	(0.08)	(0.05)	(0.07)	(0.08)	(0.07)	(0.10)	(0.08)	(0.06)
R^2	0.28	0.32	0.45	0.51	0.36	0.41	0.47	0.34	0.30
Adj. R^2	0.27	0.31	0.44	0.51	0.35	0.40	0.46	0.33	0.29
Num. obs.	122	128	126	132	124	119	103	120	121
RMSE	0.44	0.46	0.42	0.45	0.47	0.47	0.48	0.49	0.44
N Clusters	122	128	126	132	124	119	103	120	121

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Table A64. Regression of Intent to Report Headline as Harmful on Partisanship by Headline (Pro-Republican Headlines)

	DV: Intent to Report Headline as Harmful								
Headline	1	2	3	4	5	6	7	8	9
Democrat Respondent	0.44***	0.62***	0.56***	0.66***	0.69***	0.53***	0.66***	0.42***	0.45***
	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Republican Respondent	0.28***	0.35***	0.19**	0.36**	0.32***	0.38***	0.23**	0.24**	0.27**
	(0.07)	(0.08)	(0.06)	(0.10)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
R^2	0.38	0.55	0.47	0.59	0.56	0.47	0.55	0.35	0.38
Adj. R^2	0.37	0.54	0.46	0.58	0.56	0.46	0.54	0.34	0.37
Num. obs.	137	120	122	110	112	130	115	120	131
RMSE	0.47	0.49	0.44	0.48	0.46	0.50	0.43	0.46	0.47
N Clusters	137	120	122	110	112	130	115	120	131

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Perception of Headline Removal as Censorship

Table A65. Regression of Perception of Headline Removal as Censorship on Partisanship by Headline (Pro-Democrat Headlines)

	DV: Perception of Headline Removal as Censorship								
Headline	1	2	3	4	5	6	7	8	9
Democrat Respondent	0.20**	0.31***	0.43***	0.26**	0.24*	0.42***	0.22**	0.18**	0.39***
	(0.06)	(0.07)	(0.08)	(0.07)	(0.08)	(0.08)	(0.06)	(0.05)	(0.09)
Republican Respondent	0.53***	0.50***	0.77***	0.79***	0.61***	0.81***	0.60***	0.56***	0.66***
	(0.09)	(0.08)	(0.07)	(0.08)	(0.09)	(0.07)	(0.11)	(0.09)	(0.08)
R^2	0.45	0.43	0.65	0.62	0.51	0.67	0.46	0.46	0.57
Adj. R^2	0.44	0.42	0.65	0.61	0.50	0.67	0.45	0.45	0.56
Num. obs.	101	102	97	105	94	103	93	94	98
RMSE	0.47	0.48	0.47	0.42	0.46	0.45	0.43	0.45	0.49
N Clusters	101	102	97	105	94	103	93	94	98

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Table A66. Regression of Perception of Headline Removal as Censorship on Partisanship by Headline (Pro-Republican Headlines)

	DV: Perception of Headline Removal as Censorship								
Headline	1	2	3	4	5	6	7	8	9
Democrat Respondent	0.42***	0.21**	0.38***	0.26***	0.07*	0.22*	0.39***	0.28***	0.28***
	(0.08)	(0.07)	(0.08)	(0.06)	(0.03)	(0.08)	(0.09)	(0.07)	(0.07)
Republican Respondent	0.73***	0.65***	0.70***	0.55***	0.62***	0.67***	0.61***	0.67***	0.57***
	(0.07)	(0.09)	(0.08)	(0.10)	(0.09)	(0.08)	(0.09)	(0.08)	(0.09)
R^2	0.63	0.50	0.58	0.43	0.57	0.57	0.52	0.57	0.48
Adj. R^2	0.62	0.49	0.58	0.42	0.56	0.56	0.51	0.56	0.47
Num. obs.	114	104	102	93	90	92	90	94	108
RMSE	0.48	0.45	0.46	0.47	0.39	0.45	0.47	0.46	0.48
N Clusters	114	104	102	93	90	92	90	94	108

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

A2.1.7. Models With Triple Interaction Between Accuracy Question Order, Participant Partisanship, and Headline Alignment

Intent to Remove Headline

Table A67. Regression of Intent to Remove Headline on Partisanship and Alignment

	DV: Intent to Remove Headline		
	Accuracy Question First	Accuracy Question First	Accuracy Question Order Interaction
Democrat	0.76***	0.80***	0.80***
	(0.04)	(0.04)	(0.04)
Republican	0.36***	0.40***	0.40***
	(0.05)	(0.05)	(0.04)
Democrat x Pro-Democrat Headline	−0.07	−0.16**	−0.16**
	(0.06)	(0.06)	(0.06)
Republican x Pro-Republican Headline	0.03	−0.08	−0.08
	(0.07)	(0.08)	(0.07)
Accuracy Question First			−0.04
			(0.05)
Accuracy Question First x Democrat x Pro-Democrat Headline			0.09
			(0.08)
Accuracy Question First x Republican x Pro-Republican Headline			0.11
			(0.09)
R^2	0.60	0.63	0.61
Adj. R^2	0.60	0.62	0.61
Num. obs.	565	531	1096
RMSE	0.46	0.46	0.46
N Clusters	565	531	1096

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Intent to Report Headline as Harmful

Table A68. Regression of Intent to Report Headline as Harmful on Partisanship and Alignment

	DV: Intent to Report Headline as Harmful		
	Accuracy Question First	Accuracy Question First	Accuracy Question Order Interaction
Democrat	0.58***	0.68***	0.64***
	(0.05)	(0.04)	(0.04)
Republican	0.30***	0.23***	0.28***
	(0.05)	(0.04)	(0.04)
Democrat x Pro-Democrat Headline	−0.09	−0.27***	−0.23***
	(0.07)	(0.06)	(0.06)
Republican x Pro-Republican Headline	0.07	0.02	−0.02
	(0.07)	(0.07)	(0.07)
Accuracy Question First			−0.02
			(0.05)
Accuracy Question First x Democrat x Pro-Democrat Headline			0.09
			(0.08)
Accuracy Question First x Republican x Pro-Republican Headline			0.14
			(0.08)
R^2	0.46	0.49	0.47
Adj. R^2	0.46	0.49	0.47
Num. obs.	566	528	1094
RMSE	0.48	0.46	0.47
N Clusters	566	528	1094

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Perception of Headline Removal as Censorship

Table A69. Regression of Perception of Headline Removal as Censorship on Partisanship and Alignment

	DV: Perception of Headline Removal as Censorship		
	Accuracy Question First	Accuracy Question First	Accuracy Question Order Interaction
Democrat	0.28***	0.32***	0.32***
	(0.05)	(0.05)	(0.05)
Republican	0.61***	0.62***	0.63***
	(0.06)	(0.06)	(0.05)
Democrat x Pro-Democrat Headline	0.04	−0.07	−0.06
	(0.07)	(0.07)	(0.07)
Republican x Pro-Republican Headline	0.06	0.10	0.09
	(0.08)	(0.08)	(0.07)
Accuracy Question First			−0.03
			(0.05)
Accuracy Question First x Democrat x Pro-Democrat Headline			0.09
			(0.09)
Accuracy Question First x Republican x Pro-Republican Headline			−0.03
			(0.09)
R^2	0.54	0.55	0.54
Adj. R^2	0.53	0.55	0.54
Num. obs.	455	432	887
RMSE	0.46	0.46	0.46
N Clusters	455	432	887

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

A2.2. Mediation Analysis

As pre-registered, we conducted a mediation analysis that aims to test whether the effect of one variable (partisan alignment) on another (intent to remove headline, intent to report headline, perception of headline as censorship) is driven by an intermediary variable (perceived accuracy of headline). The Average Causal Mediation Effect (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.

We conduct a mediation analysis using the `mediation` package in R Tingley et al. (2014). The results are presented in Table A74, which shows that the effect of party promotion on whether a false headline should be removed is fully mediated by accuracy. We also find that the effect of political alignment on intent to report a false headline as harmful is fully mediated by accuracy. For the outcome of whether removal false content is censorship, the total effect is not significant, so although the average mediation effect is statistically significant, it is not meaningful.

The mediation analyses are based on unweighted models with standard errors that are not cluster-robust.

A2.2.1. 4-Point Accuracy Variable as Mediator

Table A70. Effect of Alignment Mediated by Accuracy for Democrats

Measure	Estimate	95% CI Lower	95% CI Upper	p-value
Intent to Remove Headline — Without Controls				
ACME	−0.065	−0.084	−0.048	< 0.001
ADE	−0.039	−0.085	0.010	0.130
Total Effect	−0.103	−0.151	−0.056	< 0.001
Proportion Mediated	0.624	0.405	1.186	< 0.001
N Observations	1302			
N Simulations	1000			
Intent to Remove Headline — With Controls				
ACME	−0.074	−0.098	−0.052	< 0.001
ADE	−0.032	−0.088	0.020	0.258
Total Effect	−0.106	−0.164	−0.050	< 0.001
Proportion Mediated	0.701	0.431	1.390	< 0.001
N Observations	995			
N Simulations	1000			
Intent to Report Headline as Harmful — Without Controls				
ACME	−0.035	−0.051	−0.023	< 0.001
ADE	−0.074	−0.127	−0.018	< 0.001
Total Effect	−0.109	−0.162	−0.055	< 0.001
Proportion Mediated	0.320	0.182	0.694	< 0.001
N Observations	1301			
N Simulations	1000			
Intent to Report Headline as Harmful — With Controls				
ACME	−0.052	−0.072	−0.034	< 0.001
ADE	−0.040	−0.106	0.020	0.218
Total Effect	−0.092	−0.156	−0.031	0.004
Proportion Mediated	0.562	0.293	1.571	0.004
N Observations	993			
N Simulations	1000			
Perception of Headline Removal as Censorship — Without Controls				
ACME	0.022	0.010	0.035	< 0.001
ADE	−0.015	−0.068	0.041	0.584
Total Effect	0.006	−0.046	0.062	0.852
Proportion Mediated	0.394	−11.203	12.449	0.852
N Observations	1032			
N Simulations	1000			
Perception of Headline Removal as Censorship — With Controls				
ACME	0.017	0.005	0.033	< 0.001
ADE	−0.004	−0.068	0.056	0.904
Total Effect	0.014	−0.049	0.073	0.660
Proportion Mediated	0.380	−7.921	7.275	0.660
N Observations	817			
N Simulations	1000			

Note: Mediation models were run with standard standard errors and without weigthing observations using a dataset in which missing values were addressed using listwise deletion.

Table A72. Effect of Alignment Mediated by Accuracy for Republicans

Measure	Estimate	95% CI Lower	95% CI Upper	p-value
Intent to Remove Headline — Without Controls				
ACME	−0.036	−0.056	−0.019	< 0.001
ADE	0.039	−0.021	0.104	0.220
Total Effect	0.003	−0.060	0.065	0.926
Proportion Mediated	−0.459	−15.548	19.612	0.926
N Observations	856			
N Simulations	1000			
Intent to Remove Headline — With Controls				
ACME	−0.048	−0.073	−0.027	< 0.001
ADE	0.053	−0.021	0.122	0.156
Total Effect	0.006	−0.067	0.077	0.902
Proportion Mediated	−0.618	−24.306	28.703	0.902
N Observations	673			
N Simulations	1000			
Intent to Report Headline as Harmful — Without Controls				
ACME	−0.018	−0.031	−0.008	< 0.001
ADE	0.044	−0.014	0.107	0.154
Total Effect	0.026	−0.033	0.088	0.422
Proportion Mediated	−0.389	−5.425	6.218	0.422
N Observations	858			
N Simulations	1000			
Intent to Report Headline as Harmful — With Controls				
ACME	−0.021	−0.036	−0.008	< 0.001
ADE	0.038	−0.033	0.104	0.264
Total Effect	0.018	−0.052	0.085	0.600
Proportion Mediated	−0.385	−11.398	9.459	0.600
N Observations	675			
N Simulations	1000			
Perception of Headline Removal as Censorship — Without Controls				
ACME	0.032	0.016	0.054	< 0.001
ADE	−0.023	−0.090	0.048	0.496
Total Effect	0.009	−0.058	0.081	0.842
Proportion Mediated	0.493	−17.209	14.184	0.842
N Observations	719			
N Simulations	1000			
Perception of Headline Removal as Censorship — With Controls				
ACME	0.039	0.020	0.065	< 0.001
ADE	−0.020	−0.101	0.056	0.606
Total Effect	0.020	−0.063	0.096	0.596
Proportion Mediated	0.763	−10.349	17.116	0.596
N Observations	572			
N Simulations	1000			

Note: Mediation models were run with standard standard errors and without weigthing observations using a dataset in which missing values were addressed using listwise deletion.

Table A74. Effect of Alignment Mediated by Accuracy

Measure	Estimate	95% CI Lower	95% CI Upper	p-value
Intent to Remove Headline — Without Controls				
ACME	−0.061	−0.076	−0.047	< 0.001
ADE	−0.001	−0.040	0.040	0.948
Total Effect	−0.062	−0.103	−0.022	< 0.001
Proportion Mediated	0.978	0.590	2.918	< 0.001
N Observations	2158			
N Simulations	1000			
Intent to Remove Headline — With Controls				
ACME	−0.061	−0.078	−0.046	< 0.001
ADE	0.002	−0.040	0.043	0.912
Total Effect	−0.060	−0.104	−0.017	0.006
Proportion Mediated	1.033	0.584	3.128	0.006
N Observations	1668			
N Simulations	1000			
Intent to Report Headline as Harmful — Without Controls				
ACME	−0.033	−0.044	−0.024	< 0.001
ADE	−0.023	−0.063	0.020	0.292
Total Effect	−0.056	−0.097	−0.014	< 0.001
Proportion Mediated	0.581	0.314	2.311	< 0.001
N Observations	2159			
N Simulations	1000			
Intent to Report Headline as Harmful — With Controls				
ACME	−0.037	−0.050	−0.024	< 0.001
ADE	−0.011	−0.055	0.033	0.646
Total Effect	−0.047	−0.092	−0.004	0.032
Proportion Mediated	0.764	0.323	3.904	0.032
N Observations	1668			
N Simulations	1000			
Perception of Headline Removal as Censorship — Without Controls				
ACME	0.038	0.026	0.053	< 0.001
ADE	−0.030	−0.075	0.017	0.214
Total Effect	0.008	−0.036	0.055	0.766
Proportion Mediated	1.033	−22.357	28.987	0.766
N Observations	1751			
N Simulations	1000			
Perception of Headline Removal as Censorship — With Controls				
ACME	0.027	0.017	0.039	< 0.001
ADE	−0.012	−0.058	0.034	0.634
Total Effect	0.015	−0.031	0.058	0.546
Proportion Mediated	0.819	−16.090	16.609	0.546
N Observations	1389			
N Simulations	1000			

Note: Mediation models were run with standard standard errors and without weigthing observations using a dataset in which missing values were addressed using listwise deletion. Models with controls include participant partisanship as a control.

A2.2.2. Binary Accuracy Variable as Mediator

Table A76. Effect of Alignment Mediated by Accuracy for Democrats

Measure	Estimate	95% CI Lower	95% CI Upper	p-value
Intent to Remove Headline — Without Controls				
ACME	−0.049	−0.065	−0.034	< 0.001
ADE	−0.055	−0.101	−0.007	0.022
Total Effect	−0.103	−0.151	−0.056	< 0.001
Proportion Mediated	0.470	0.295	0.880	< 0.001
N Observations	1302			
N Simulations	1000			
Intent to Remove Headline — With Controls				
ACME	−0.056	−0.078	−0.036	< 0.001
ADE	−0.050	−0.106	0.003	0.060
Total Effect	−0.106	−0.163	−0.050	< 0.001
Proportion Mediated	0.532	0.316	1.064	< 0.001
N Observations	995			
N Simulations	1000			
Intent to Report Headline as Harmful — Without Controls				
ACME	−0.019	−0.030	−0.009	< 0.001
ADE	−0.091	−0.144	−0.036	< 0.001
Total Effect	−0.109	−0.162	−0.055	< 0.001
Proportion Mediated	0.168	0.077	0.399	< 0.001
N Observations	1301			
N Simulations	1000			
Intent to Report Headline as Harmful — With Controls				
ACME	−0.029	−0.045	−0.015	< 0.001
ADE	−0.063	−0.129	−0.001	0.044
Total Effect	−0.092	−0.157	−0.030	0.004
Proportion Mediated	0.316	0.144	0.912	0.004
N Observations	993			
N Simulations	1000			
Perception of Headline Removal as Censorship — Without Controls				
ACME	0.017	0.008	0.029	< 0.001
ADE	−0.011	−0.064	0.044	0.700
Total Effect	0.006	−0.046	0.062	0.850
Proportion Mediated	0.340	−9.627	9.808	0.850
N Observations	1032			
N Simulations	1000			
Perception of Headline Removal as Censorship — With Controls				
ACME	0.015	0.005	0.027	< 0.001
ADE	−0.001	−0.065	0.059	0.978
Total Effect	0.014	−0.049	0.072	0.650
Proportion Mediated	0.333	−7.630	5.448	0.650
N Observations	817			
N Simulations	1000			

Note: Mediation models were run with standard standard errors and without weigthing observations using a dataset in which missing values were addressed using listwise deletion.

Table A78. Effect of Alignment Mediated by Accuracy for Republicans

Measure	Estimate	95% CI Lower	95% CI Upper	p-value
Intent to Remove Headline — Without Controls				
ACME	−0.030	−0.047	−0.014	< 0.001
ADE	0.033	−0.029	0.096	0.318
Total Effect	0.003	−0.058	0.066	0.938
Proportion Mediated	−0.341	−11.283	20.538	0.938
N Observations	856			
N Simulations	1000			
Intent to Remove Headline — With Controls				
ACME	−0.039	−0.061	−0.020	< 0.001
ADE	0.044	−0.031	0.114	0.236
Total Effect	0.006	−0.067	0.076	0.878
Proportion Mediated	−0.523	−24.742	15.658	0.878
N Observations	673			
N Simulations	1000			
Intent to Report Headline as Harmful — Without Controls				
ACME	−0.015	−0.027	−0.006	< 0.001
ADE	0.042	−0.018	0.103	0.170
Total Effect	0.026	−0.032	0.088	0.414
Proportion Mediated	−0.332	−5.783	5.737	0.414
N Observations	858			
N Simulations	1000			
Intent to Report Headline as Harmful — With Controls				
ACME	−0.018	−0.033	−0.006	< 0.001
ADE	0.036	−0.036	0.102	0.290
Total Effect	0.018	−0.052	0.086	0.600
Proportion Mediated	−0.331	−9.789	9.286	0.600
N Observations	675			
N Simulations	1000			
Perception of Headline Removal as Censorship — Without Controls				
ACME	0.029	0.013	0.047	< 0.001
ADE	−0.020	−0.088	0.050	0.568
Total Effect	0.009	−0.059	0.080	0.836
Proportion Mediated	0.455	−13.882	10.481	0.836
N Observations	719			
N Simulations	1000			
Perception of Headline Removal as Censorship — With Controls				
ACME	0.036	0.017	0.060	< 0.001
ADE	−0.016	−0.099	0.060	0.670
Total Effect	0.020	−0.061	0.096	0.594
Proportion Mediated	0.690	−7.838	14.338	0.594
N Observations	572			
N Simulations	1000			

Note: Mediation models were run with standard standard errors and without weigthing observations using a dataset in which missing values were addressed using listwise deletion.

Table A80. Effect of Alignment Mediated by Accuracy

Measure	Estimate	95% CI Lower	95% CI Upper	p-value
Intent to Remove Headline — Without Controls				
ACME	−0.045	−0.058	−0.033	< 0.001
ADE	−0.017	−0.056	0.024	0.430
Total Effect	−0.062	−0.102	−0.021	< 0.001
Proportion Mediated	0.718	0.422	2.136	< 0.001
N Observations	2158			
N Simulations	1000			
Intent to Remove Headline — With Controls				
ACME	−0.047	−0.062	−0.033	< 0.001
ADE	−0.013	−0.055	0.028	0.556
Total Effect	−0.060	−0.103	−0.018	0.006
Proportion Mediated	0.792	0.422	2.281	0.006
N Observations	1668			
N Simulations	1000			
Intent to Report Headline as Harmful — Without Controls				
ACME	−0.021	−0.029	−0.013	< 0.001
ADE	−0.035	−0.076	0.007	0.116
Total Effect	−0.056	−0.096	−0.014	< 0.001
Proportion Mediated	0.362	0.186	1.438	< 0.001
N Observations	2159			
N Simulations	1000			
Intent to Report Headline as Harmful — With Controls				
ACME	−0.023	−0.034	−0.013	< 0.001
ADE	−0.024	−0.069	0.020	0.302
Total Effect	−0.047	−0.092	−0.005	0.028
Proportion Mediated	0.495	0.189	2.817	0.028
N Observations	1668			
N Simulations	1000			
Perception of Headline Removal as Censorship — Without Controls				
ACME	0.029	0.018	0.041	< 0.001
ADE	−0.022	−0.066	0.025	0.368
Total Effect	0.008	−0.037	0.055	0.774
Proportion Mediated	0.790	−22.352	19.401	0.774
N Observations	1751			
N Simulations	1000			
Perception of Headline Removal as Censorship — With Controls				
ACME	0.022	0.014	0.033	< 0.001
ADE	−0.008	−0.054	0.038	0.752
Total Effect	0.015	−0.032	0.058	0.552
Proportion Mediated	0.678	−12.219	12.368	0.552
N Observations	1389			
N Simulations	1000			

Note: Mediation models were run with standard standard errors and without weigthing observations using a dataset in which missing values were addressed using listwise deletion. Models with controls include participant partisanship as a control.

References

Allcott, Hunt and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31(2):211–236.
URL: *https://doi.org/10.1257/jep.31.2.211 https://www.aeaweb.org/content/file?id=4222*

Honaker, J, G King and M Blackwell. 2011. "Amelia II: A program for missing data." *Journal of Statistical Software* 45(7):1–3.

Mosleh, M., C. Martel, Dean Eckles and David G. Rand. 2021. 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 *ACM Conference on Human Factors in Computing Systems (CHI '21)*.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele and Kosuke Imai. 2014. "mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software* 59(5):1–38.
URL: *http://www.jstatsoft.org/v59/i05/*