$9^{th}$ International Conference on Computational Social Science $IC^2S^2$
July 17-20, 2023, Copenhagen, Denmark

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

*Keywords: social media, partisanship, censorship, content moderation, misinformation*

## Extended Abstract

Misinformation is seen as a major global threat by political and economic leaders around the world and the general public [7]. Rising public awareness of online misinformation has coincided with growing public debates about what social media companies should remove from their platforms. These debates have laid bare deep partisan divisions over the removal of online content in the United States [5], which have stymied efforts to deal with misinformation.

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 [1], 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 pre-registered survey of 1,120 U.S. respondents where we present participants misinformation headlines and where we vary the partisan alignment of the misinformation headlines (see Figure 4). 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 control for the "fact gap" by explicitly telling respondents that the headlines are misinformation, and, for some analyses, by subsetting to respondents who agree with this assessment. We also analyze the relationship between partisanship and accuracy, and whether accuracy mediates the effects of partisanship on content moderation preferences.

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. Figure 1 plots the coefficient estimates and confidence interval for our main model, an OLS regression interacting partisanship of participants and political alignment of the headlines for all respondents. It shows large overall differences in content moderation preferences, regardless of headline alignment.

While we find some evidence of Democrats' willingness to use content moderation for party promotion (see third row of Figure 1), 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.

Both Democrats and Republicans are more likely to think that headlines that align with their own position are true, reflecting a fact gap (see Figure 2). These different accuracy perceptions do not fully mediate the effect of partisanship on content moderation preferences, however. Our main findings, a strong value gap and limited party promotion, hold both when including all participants (Figure 1) or only participants that rated the headlines as inaccurate (Figure 3).

The results of this experiment show that the value gap plays a huge role in attitudes toward removal of misinformation online. This 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 [3]. An alternative explanation could be elite signaling. Previous research has shown that elite signaling can drive opinions [6], and Republican elites have signaled their opposition to deplatforming and extensive content moderation, while Democrats have supported it.

Further research is needed to explore how these findings generalize from a survey experiment to social media platforms. In our experiment, we focused on internal validity and balanced the partisanship of headlines and kept other headline characteristics and their context relatively comparable. Future research should investigate further how specific types of content and different contexts influence the value gap and party promotion.

These findings have important implications for policymakers. In an environment with increasing partisan animosity, it is encouraging that the effects of party promotion are dwarfed by the value gap given that respondents – Republicans in particular – seemed to evaluate content removal outside of the lens of party promotion. However, the results also suggest that settling factual disagreements will not resolve partisan conflict over content moderation.

Policymakers and social media platforms could consider different approaches to design policies with bipartisan support, including focusing on agreement on the system of moderation procedures rather than on specific pieces of content [2], using less extreme forms of content moderation like flagging or downranking misinformation, or using moral reframing [4] 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

[1] Larry M. Bartels. *Unequal Democracy*. Princeton University Press, Princeton, NJ, 2016.

[2] Evelyn Douek. Content Moderation as Systems Thinking. *Harvard Law Review*, 136, 2022.

[3] Matthew Feinberg and Robb Willer. From Gulf to Bridge: When Do Moral Arguments Facilitate Political Influence? *Personality and Social Psychology Bulletin*, 41(12):1665–1681, 2015.

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

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

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

[7] Pew Research Center. Climate Change Remains Top Global Threat Across 19-Country Survey. Technical report, 2022.

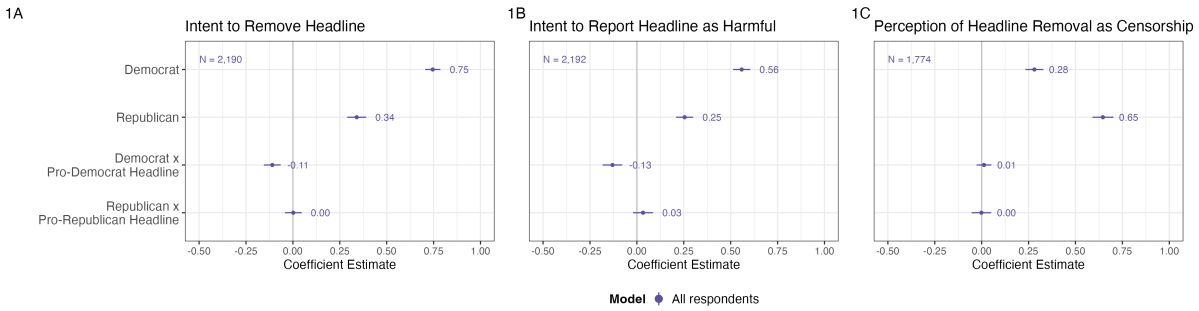


Figure 1: Partisanship and preferences for content moderation for all respondents (models without control variables).

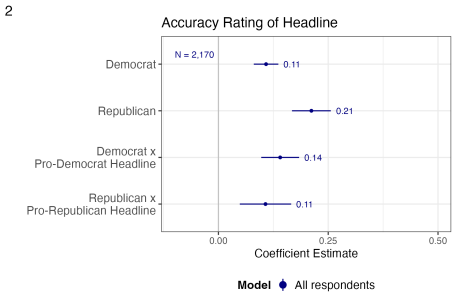


Figure 2: Respondents' assessment of headline accuracy (models without control variables). Model does not include an intercept term.

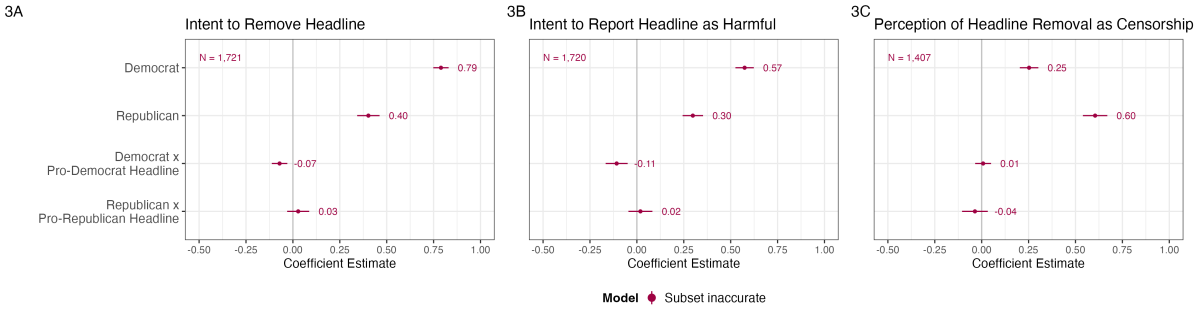


Figure 3: Partisanship and preferences for content moderation for respondents who agree that headlines are inaccurate (models without control variables).

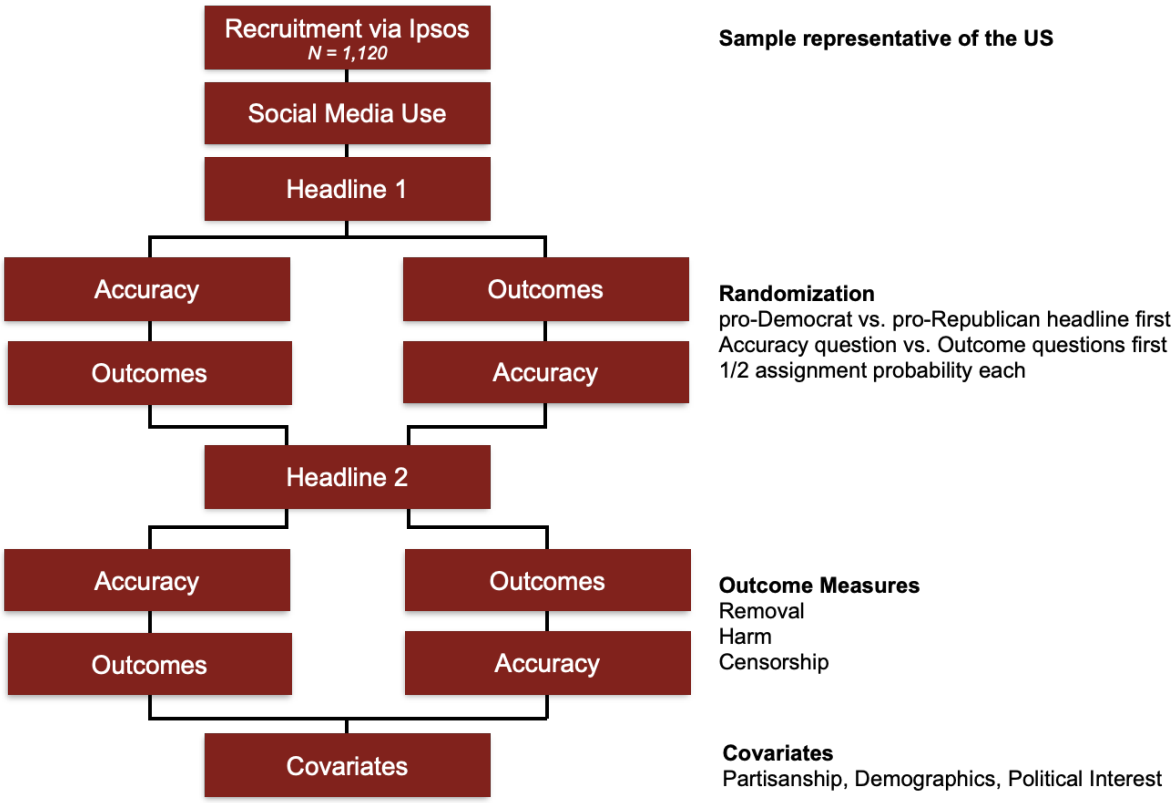


Figure 4: Experiment design overview