Moral Alignment Shapes Responses to Shared Content

Suhaib Abdurahman[1,3], Nils K. Reimer[4], Preni Golazizian[2], Elisa Baek[1], Yixuan Shen[5], Jackson Trager[1,3], Roshni Lulla[1,3], Jonas Kaplan[1,3], Carolyn Parkinson[5], and Morteza Dehghani[1,2,3]

[1]Department of Psychology, University of Southern California

[2]Department of Computer Science, University of Southern California

[3]Brain and Creativity Institute, University of Southern California

[4]Department of Psychological & Brain Sciences, University of California, Santa Barbara

[5]Department of Psychology, University of California, Los Angeles

Author Note

Suhaib Abdurahman https://orcid.org/0000-0001-5615-0129

All data, analysis code, and research materials, and preregistrations are available under https://osf.io/z25tc/?view_only=0141845d12024a2cbdbd0f71f77f23a8. The authors declare no conflicts of interest. The project received funding from NSF SaTC (Award No: 2140473).

Correspondence regarding this article should be addressed to Suhaib Abdurahman, Department of Psychology, University of Southern California, SGM 501, 3620 S. McClintock Ave., Los Angeles, CA 90089-1061. Email: sabdurah@usc.edu

Abstract

Does targeting audiences' core values facilitate the spread of misinformation? We investigate this question by analyzing real-world Twitter data (N = 20,235; 809,414 tweets) on misinformation regarding COVID vaccinations, in conjunction with a set of behavioral experiments. First, we use natural language processing to determine messages' moral framing and stance on COVID vaccinations and mandates and find that an alignment of moral framing and stance (e.g., Binding framing and "anti-vax") facilitates the spread of COVID vaccination misinformation. We then replicate our findings in three behavioral experiments, two of which were preregistered ($N_{2a} = 615$; $N_{2b} = 505$; $N_3 = 533$). We investigate how the effect of aligning messages' moral framing with participants' moral values impacts participants intentions to share true and false news headlines and whether this effect is driven by a lack of analytical thinking. Our results show that framing a post such that it aligns with audiences' moral values leads to increased sharing intentions, independent of headline veracity, headline familiarity, and participants' political ideology. However, we find no effect of analytical thinking on misinformation sharing and plausibility concerns. Our findings suggest that (a) targeting audiences' core values can be used to influence the dissemination of (mis)information on social media platforms, (b) partisan divides in misinformation sharing can be, at least partially, explained through alignment between audiences' underlying moral values and moral framing that often accompanies content shared online, and (c) this effect is driven by motivational factors and not lack of deliberation.

Keywords: misinformation, fake news, moral values, moral foundations theory, social media, natural language processing, information sharing

Moral Alignment Shapes Responses to Shared Content

The prevalence of misinformation poses an imminent threat to our society. Increasingly, cyberattacks leverage social media networks to malevolently influence audiences, undermining civil discourse by instigating division and polarization (Allcott & Gentzkow, 2017; Grinberg et al., 2019; Guess et al., 2018; Lazer et al., 2017; Nyilasy, 2019; Vosoughi et al., 2018). Most notably, malicious actors have manipulated narratives, amplified inflammatory messages, anda distorted public opinion, as highlighted by The US Senate Investigation Committee on Russian Interference into the 2016 US Election and the January 6^{th} committee (Bossetta, 2018; Jensen, 2018; Mueller III, 2019; Yin et al., 2018; Ziegler, 2018). Similar adversarial operations have been documented in other democratic countries all over the world, such as during the Brexit campaign in the UK or elections in Brazil and India (Aral & Eckles, 2019). The scope and severity of these attacks make it important to identify the specific psychological strategies used by malicious actors to spread targeted misinformation in order to mitigate vulnerabilities to such attacks. This requires an understanding of the social and cognitive mechanisms underlying how such attacks impact audiences.

Past research on misinformation has identified cognitive, affective, and social factors that drive the belief in, and spread of, misinformation. Cognitive heuristics and peripheral cues such as familiarity, processing fluency and cohesion have been found to increase acceptance of misinformation (Ecker et al., 2022; Schwarz et al., 2016), independent of ability and prior knowledge (De Keersmaecker et al., 2020; Fazio, 2020; Fazio et al., 2015). Affective factors, such as mood and emotions, have been linked to susceptibility to misinformation through increased reliance on processing fluency and decreased skepticism (Forgas & East, 2008; Koch & Forgas, 2012; Martel et al., 2020). Social factors, such as perceived source credibility, have been found to affect belief in misinformation and people are generally more likely to trust sources that are aligned with their values and worldview (Brinol & Petty, 2009; Ecker et al., 2022; Mackie et al., 1990; Mahmoodi et al., 2015).

A large body of literature further points to the role of prior beliefs in sharing and believing misinformation through motivated reasoning (Kunda, 1990; Taber & Lodge, 2006). Misinformation that aligns with one's moral and political attitudes is perceived as more accurate and reliable (Ecker et al., 2022; Van Bavel & Pereira, 2018; Winkielman et al., 2012) and readers tend to share or leave positive comments on content that resonates with their political beliefs (Colliander, 2019; Pennycook & Rand, 2019c). Furthermore, past research has shown that using moral-emotional language generally increases the virality and spread of messages on social media platforms, due to increased attention (Brady et al., 2020; Brady et al., 2017; Valenzuela et al., 2017) and resonance with audiences (Adger et al., 2017; Hurst & Stern, 2020). This indicates that misinformation campaigns could utilize moral language not only to persuade users but also to achieve extensive spread in these networks and thus reach a vast number of users. However, focusing on the mere presence of moral language is too simplistic of an approach to explain differences in behavior relating to misinformation. Some studies observe interactions effects between specific kinds of moral language and person-level variables, such as ideology (Erceg et al., 2018; Kivikangas et al., 2021; Low et al., 2016) and other demographics (Kivikangas et al., 2021). Some research even observed "backfire" effects, that is, a reduction in engagement and online diffusion, based on the frequency and type of moral language in messages (Candia et al., 2022).

Related to the study of (moral) language used in messages shared online, framing effects have been discussed in past research on judgments and behaviors regarding moral and political issues (Hoover et al., 2018; Sunstein, 2003). Specifically, moral framing can lead to persuasion even in highly partisan settings, such that political arguments that are framed in line with audiences' moral concerns are more successful in persuading audiences (Day et al., 2014; Feinberg & Willer, 2019; Voelkel et al., 2022). For instance, framing pro-environmental behavior in moral language relating to values typically endorsed by conservatives (i.e., loyalty, purity, patriotism or duty), promoted pro-environmental

attitudes among conservatives even though these behaviors are typically associated with liberals (Feinberg & Willer, 2013, 2015; Wolsko et al., 2016). Importantly, the specific language and framing used influenced the acceptance of information beyond political beliefs conveyed in the very same message (i.e., the message being pro-Democrat or pro-Republican). This suggests that modern misinformation campaigns can gain efficacy in part by "personalizing" claims to resonate with the core moral concerns of their intended audiences and gain legitimacy and strength from the value-laden and moral claims they make. The use of moral framing can also lead to the moralization or sacrilization of issues (Marietta, 2008) which in turn influences group behavior and attitudes, such as increasing polarization and inciting outrage and violence against outgroups (Dehghani et al., 2010; Graham & Haidt, 2012). Specifically, the moralization of issues can activate moral convictions which are linked to rigid, absolutist mindsets (Skitka et al., 2005) and thus an overt focus on achieving morally mandated goals (Skitka & Mullen, 2002) by potentially engaging in and justifying extreme actions (Skitka et al., 2005; Skitka & Morgan, 2014; Skitka & Mullen, 2002). Therefore, it is critical to understand how misinformation campaigns use moral language in order to mitigate these severe consequences.

Our work, thus, seeks to elucidate how misinformation campaigns can use moral framing to effectively persuade their audiences and how this strategy plays out in the real world. Specifically, this work investigates the effect of matching message framing and individuals' values on the spread of targeted (mis)information. Our work relies on the Moral Foundations Theory (MFT; Graham, 2013), an intuition-driven pluralistic model of morality, to operationalize individuals' moral values. In this model, moral values are composed of two superordinate, bipolar categories (Atari et al., 2020; Graham, 2013; Graham & Haidt, 2010; Haidt & Graham, 2007; Haidt et al., 2009; Haidt & Joseph, 2004): Individualizing (i.e., focused on individuals' rights and well-being) and Binding values (i.e., focused on group preservation)[1]. This more specific and granular perspective on both the

[1] Note, that recent research suggests that these superordinate categories might be specific to Western

message content and individuals' values provides additional nuances to the psychological drivers of misinformation and the role of morality in people's decision-making in regard to information sharing. Adopting the MFT framework, our work adds to past literature that only investigated the general presence of moral language in shared content (Brady et al., 2020; Brady et al., 2017; Valenzuela et al., 2017) or the impact of aligning the content of misinformation and audience worldview on acceptance and spread of misinformation (Colliander, 2019; Ecker et al., 2022; Pennycook & Rand, 2019c; Van Bavel & Pereira, 2018; Winkielman et al., 2012).

We hypothesize that messages that align with audiences' core moral values, independent of being true or false, will be more effective than those which are misaligned or which do not target core moral values. We expect that, in the U.S., misinformation campaigns that rely on moral framing centered around Binding values are more effective in specifically persuading political conservatives, and conversely, that misinformation campaigns that rely on Individualizing framing are more effective in specifically persuading liberals to believe and share misinformation. Our hypotheses are based on the observation that, across countries and cultures, liberals tend to prioritize Individualizing values instead of Binding values, while conservatives value Individualizing and Binding values more equally (Graham et al., 2009). A recent meta-analysis of 89 samples and 226,674 participants found that Individualizing values correlate negatively and Binding values correlate positively with political conservatism (Kivikangas et al., 2021).

Further, we test the hypothesis that the proposed effects of moral values and framing might be driven by a lack of deliberation. Previous work has argued that "analytical thinking", and more generally trait-level deliberation tendency, reduces belief in and sharing of misinformation (Pennycook & Rand, 2019c, 2021) and that moral language increases the spread of messages via increased attention capture (Brady et al., 2020). In line with the classical reasoning approach, which suggests that people share misinformation

---

cultures (Atari et al., 2020)

because they do not notice it is misinformation ("lack of deliberate thinking"), it could be that aligned moral framing distracts participants from deliberating over sharing a post and thus from the shared information being false or implausible. If true, then the effect of aligning moral values and message framing should be mediated by deliberating over sharing a post. Alternatively, participants could be motivated by their intuitions of right and wrong that accompany moralized posts (see work on motivated reasoning and specifically how moral values motivate behavior: Dehghani et al. (2016) and Kahan et al. (2017)) and that these intuitions supersede accuracy concerns. In that case, there should not be an effect of deliberation, both trait-level and measured for each post, on sharing of (mis)information.

To investigate the relationship between moral framing and responses to shared content, we conduct two sets of studies. First, we analyze real-world social media (Twitter) conversations about COVID-19 vaccinations and mandates regarding the relationship between a message's moral framing and the sender's stance on this issue. Second, we develop a paradigm that allows us to directly test how the specific match of moral framing and audiences' moral values affect responses to shared social media content in a controlled experimental paradigm. We then use this paradigm in two pre-registered studies to confirm the proposed effects and to shed light on the underlying psychological mechanisms. Together, our work provides additional insight into how the specific alignment of moral values and message framing may contribute to the spread of (mis)information.

Significance Statement

The spread of misinformation has become a major concern to society, particularly in the age of social media. We show here that aligning online messages with audiences' core moral values leads to increased sharing, independent of message veracity, message familiarity, and users' analytical thinking ability. The results suggest that misinformation is driven by motivational factors, such as alignment with one's core moral values, above and beyond cognitive factors. Our findings further indicate the susceptibility of audiences

to manipulation via simple framing strategies, their potential use in targeted misinformation campaigns, and thus the need to develop effective countermeasures.

Study 1

In Study 1, we analyze COVID-related content on Twitter regarding the relationship between tweets' moral framing, users' stance on the topic, and liking or sharing of the tweets. We predict that moral framing that matches values associated with a stance (e.g., liberal and Individualizing values) will lead to increased sharing and liking of tweets. Previous research has documented that stance on COVID-19 vaccinations is strongly related to political ideology (Clarkson & Jasper, 2022; Jiang et al., 2021; Kerr et al., 2021; Stroope et al., 2021) with more conservatives endorsing anti-vaccination ("anti-vax") attitudes and more liberals endorsing pro-vaccination ("pro-vax") attitudes. Since Individualizing values correlate negatively and Binding values correlate positively with political conservatism (see Kivikangas et al., 2021), we expect that content by "anti-vax" users would be shared more frequently, compared to content by "pro-vax" users, when framed with Binding values. Conversely, we expect that content by "pro-vax" users would be shared more frequently, compared to content by "anti-vax" users, when framed with Individualizing values. We also expect to replicate previous findings of liberals prioritizing Individualizing over Binding values and conservatives endorsing both equally (Graham et al., 2009). This means that we predict Individualizing framing to be more effective than Binding framing for content by "pro-vax" users, but both framing to be equally effective for content by "anti-vax". Note, that this is a within-group comparison (i.e., within "pro-vax" and within "anti-vax"), whereas the previous hypotheses were between-group comparisons (i.e., between "pro-vax" and "anti-vax").

Method

We collected social media messages about COVID vaccinations and mandates from Twitter and used state-of-the-art natural language processing methods to extract the

messages' moral framing and users' stance on this issue. Finally, we fit a model predicting liking and sharing of these messages as a function of messages' moral framing, users' stance on COVID vaccinations, and their interaction.

Data Collection. We utilized an existing corpus of tweets, specifically rumors and misinformation, on COVID-19 vaccinations and mandates compiled by Muric et al. (2021). We collected a random sample of 809,414 tweets spanning from June 2021 to November 2021 (most current tweets at the time of data collection) using the Twitter API. Other than the tweet text, we collected meta-data, including the user-id, dates, number of retweets, and favorite count (i.e., "likes").

Procedure. We used a Bidirectional Encoder Representations from Transformers (BERT)-based (Devlin et al., 2018) classifier to determine the moral language in each tweet with the tweet text as input. Specifically, we used the pre-trained BERT model "small BERT" (Turc et al., 2019) with $L = 12$ hidden layers (i.e., Transformer blocks), a hidden size of $H = 256$, and $A = 4$ attention heads. We added a downstream classification layer to the language model to predict whether a tweet contained moral vs. non-moral language, and for the moral messages whether these were framed using Individualizing or Binding foundations. We simultaneously trained the classification layer and fine-tuned the embedding layers on the Moral Foundations Twitter Corpus (Hoover et al., 2020), which is an annotated corpus containing 35,108 tweets along with each tweet's moral framing based on the Moral Foundations framework (Graham et al., 2013). The classifier achieved a cross-validated F_1 score of 0.84 for moral/non-moral message classification and 0.76 when predicting Binding vs. Individualizing framing.

We further inferred each user's position on COVID vaccination and mandates. More specifically, we employed an unsupervised stance detection method (Darwish et al., 2020) which uses dimensionality reduction to project users onto a low-dimensional space, followed by clustering, that allows identifying representative core users. To classify the stance of each user in the corpus as either pro-vaccination ("pro-vax") or anti-vaccination

("anti-vax"), we compute the cosine similarity between each pair of users based on (1) (re-)tweeting identical tweets; (2) the hashtags that users use; and (3) the accounts they retweet. We then manually checked a random sample of 1000 users by evaluating the tweets and keywords they posted or retweeted (36,026), and based on manual verification, the stance of 85% of users had been classified correctly.

Measures. In our final data set, each tweet, in addition to the number of retweets and "likes", had the following additional information associated with it[2]:

- Moral Framing: Whether it contained moral or non-moral language.
- Binding & Individualizing framing: Whether the moral messages were framed using Binding and/or Individualizing or non-moral language.
- Stance: Whether the tweet comes from a user who is "pro-vax" or "anti-vax".

In total, 64% of tweets were posted by "anti-vax" users (vs. 36% by "pro-vax" users), 25% of tweets contained moral framing (vs. 75% non-moral framing), 6% of tweets containing Binding framing and 18% Individualizing framing.

Analysis Strategy. Focusing on our hypothesis, we analyzed our data to determine whether people engage more (measured via the number of retweets and favorites) with a social media post if the framing of the post aligned with the values associated with the posts' stance (e.g., Binding values with "anti-vax" posts). Specifically, we ran a series of negative binomial models that predicted the number of retweets or likes (separate outcome variables) as a function of various predictor variables. Model 0 estimated the number of likes as a function of the user's stance ("pro-vax" vs. "anti-vax") and included a fixed intercept and a varying (random) intercept accounting for variance across users, with the average number of tweets per user being 40. Model 1 extended Model 0 by estimating the number of likes as a function of a tweet's moral framing (Individualizing and Binding) and including a random effect accounting for variance in

[2] See A1 for example messages covering the different framing and stances.

framing effects over users. Model 2 extended Model 1 by estimating the number of likes as a function of the interaction between a tweet's moral framing and the user's stance. Thus, Model 2 functioned as our main model and tested our hypotheses of moral framing and stance interactions being predictive of message engagement while showing the specific underlying dynamic (e.g, the effect of individualizing vs. binding framing on likes for pro-vax tweets). We further ran the same series of models with the number of retweets as an alternative outcome variable for user engagement.

To estimate these models, we used the 'brms' R package (Version 2.16.1) (Bürkner, 2017, 2018) as an interface to fit Bayesian generalized linear multilevel models in Stan (Stan Development Team, 2021). Bayesian inference involves choosing a likelihood function and prior distributions. The likelihood function links the observed data to one or more model parameters (e.g., regression coefficients) by expressing how likely the observed data would have been for different values of said model parameters. Prior distributions state how plausible different values of said model parameters are before considering the observed data. Our models used weakly informative prior distributions, Student-t$(3, 0, 2.5)$, for all model parameters. Bayesian inference applies Bayes' theorem to update prior distributions in light of the observed data to produce posterior distributions. Posterior distributions state how plausible different values of the model parameters are given the observed data. We report point estimates, based on the median of posterior samples, and 95% uncertainty intervals, based on the quantiles of posterior samples, for relevant model parameters.

We used 10-fold cross-validation to compare how well each model predicted sharing intentions outside the sample used to estimate it. As a measure of out-of-sample prediction accuracy, we calculated each model's expected log predictive density ($ELPD$), that is, the logarithm of the joint posterior predictive probability of all observations. To compare models, we calculated the difference in out-of-sample prediction accuracy for each pair of models (Δ_{ELPD}), with positive values indicating that a model made more accurate predictions than a comparison model (Vehtari et al., 2017). We divided this difference by

its standard error ($z = \Delta_{ELPD}/SE$) to account for the uncertainty of cross-validation as an estimate of out-of-sample prediction accuracy. We selected a more complex over a simpler model when the difference in prediction accuracy was at least 1.96 times larger than its standard error.[3]

Results

Table 1 compares each model's out-of-sample prediction accuracy of engagement, captured by retweet count to that of the null model without predictors (M0) and that of the other models with predictors (M1–M2). We found that Model 2 —which included tweet's moral framing (Binding and Individualizing) and their interactions with the user's stance ("pro-vax" and "anti-vax")— predicted engagement more accurately than Model 0 ($\Delta_{ELPD} = 59.8, SE = 9.8, z = 6.10$) and Model 1 ($\Delta_{ELPD} = 25.7, SE = 6.1, z = 4.23$), indicating the relevance of matching moral framing and individuals' values for the spread of social media messages. The between-group analyses shows, as hypothesized, that tweets' Individualizing framing predicted more (1.6 times) engagement when posted by "pro-vax" users compared to "anti-vax" users ($\beta = 0.20, [0.09, 0.31]$). Conversely, Binding framing predicted more (1.7 times) engagement when posted by "anti-vax" users compared to "pro-vax" users ($\beta = -0.23, [-0.42, -0.03]$). The within-group analyses show, as hypothesized, that Individualizing framing predicted significantly more engagement (2.6 times) than Binding framing for "pro-vax" users ($\beta = 0.41, [0.21, 0.60]$), as well as no difference between both framing for "anti-vax" users ($\beta = -0.02, [-0.15, 0.11]$).

Analogously to Table 1, Table 2 compares each model's out-of-sample prediction accuracy of engagement, captured by favorite count, to that of the null model without predictors (M0) and that of the other models with predictors (M1–M2). Supporting Hypothesis 1, Model 2 —that included tweets' moral framing (Binding and Individualizing)

[3] For a conventional interpretation, consider the critical values to be $|\Delta_{ELPD}/SE| > 1.96$ for $p < .05$; 2.58 for $p < .01$; and 3.29 for $p < .001$ in a two-sided null-hypothesis significance test of the difference in out-of-sample prediction accuracy.

Table 1
Comparison of models estimating engagement (retweet count) as a function of various predictor variables

			z		
Model	Description	R^2	M0	M1	M2
M0	Stance	0.10	-	-3.63	-6.1
M1	Moral framing	0.14	3.63	-	-4.23
M2	Moral framing & stance interaction	0.14	6.1	4.23	-

Note. R^2 *is a Bayesian analogue to the proportion of within-sample variance explained by a model (not considering varying effects).* z *is the difference in out-of-sample prediction accuracy between two models divided by its standard error* ($z = \Delta_{ELPD}/SE$).

and their interactions with the user's stance ("pro-vax" and "anti-vax")— predicted engagement more accurately than Model 0 ($\Delta_{ELPD} = 65.9, SE = 9.2, z = 7.16$) and Model 1 ($\Delta_{ELPD} = 20.6, SE = 6.7, z = 3.07$). The between-group analyses show, again as hypothesized, that tweets' Individualizing framing predicted more (12 times) engagement when posted by "pro-vax" users compared to "anti-vax" users ($\beta^4 = 1.08, [0.78, 1.38]$). However, against our hypothesis, the effect of Binding framing did not differ for "pro-vax" and "anti-vax" users ($\beta = -0.10, [-0.60, 0.41]$). The within-group analyses show that, as hypothesized, Individualizing framing predicted more engagement (6.3 times) than Binding framing for "pro-vax" users ($\beta = 0.80, [0.20, 1.36]$). However, against our hypothesis, Binding framing predicted more engagement (2.5 times) compared to Individualizing framing for "anti-vax" users ($\beta = -0.39, [-0.70, -0.06]$).

Overall, these findings show that, generally, an alignment of moral framing and stance increases engagement with social media messages. Specifically, we found that Individualizing framing facilitated engagement for "pro-vax" (compared to "anti-vax") tweets while Binding framing facilitated engagement for "anti-vax" (compared to "pro-vax") tweets. Furthermore, these results were found across both engagement metrics (i.e., liking and retweeting) with only one exception: Binding framing showed no difference

[4] Note that for negative binomial regression the regression coefficient expresses the difference in the *log* of expected outcome count for one unit change of the predictor variable.

Table 2
Comparison of models estimating engagement (favourites count) as a function of various predictor variables

			z		
Model	Description	R^2	M0	M1	M2
M0	Stance	0.047	-	-4.67	-7.16
M1	Moral framing	0.075	4.67	-	-3.07
M2	Moral framing & stance interaction	0.076	7.16	3.07	-

Note. R^2 *is a Bayesian analogue to the proportion of within-sample variance explained by a model (not considering varying effects).* z *is the difference in out-of-sample prediction accuracy between two models divided by its standard error* ($z = \Delta_{ELPD}/SE$).

between both groups regarding likes (but not retweets). This could be caused by user behavior differing for liking and sharing tweets. For example, users could be less hesitant to like content that they would not share because it is less public. Furthermore, we are not able to directly measure the political and moral values of the users and instead use their stance (pro-vax vs. anti-vax) as a proxy. While attitudes towards COVID vaccinations are indeed strongly polarized (most pro-vaxers being liberal and most anti-vaxers being conservative; Cheng et al. (2020)), there is a significant number of conservatives who are not "anti-vax". These "pro-vax" conservatives could engage with "pro-vax" messages that have a Binding framing. We therefore address these limitations in Study 2 and Study 3, which directly investigate the relationship between moral framing of messages, individual's moral values, and responses to shared social media content to explore the underlying mechanisms of information sharing in a controlled experimental paradigm.

Study 2

In this study, we conduct two behavioral experiments to (1) develop a paradigm for studying how moral framing affects responses to shared social media content (Study 2a) and (2) use this paradigm to test our hypotheses about the relationships between moral framing, moral values, and responses to shared social media content (Study 2b). Study 2

was designed to, first, confirm that matching moral framing and moral values increase liking and sharing of shared online content– in a controlled experimental paradigm and to, second, shed light on the underlying mechanisms that drive engagement with information shared online. Note that whereas Study 1 focused on the apparent moral values of message sources, Study 2 focuses on the moral values of message recipients. However, given that people tend to expose themselves to social media content that agrees with their worldview (Aiello et al., 2012; Bakshy et al., 2015) and moral values (Dehghani et al., 2016; Singh et al., 2021), it is very likely that audience engagement (favorite and retweet count) measured in Study 1 were captured from users whose moral values matched those of the relevant message source. Nevertheless, Study 2 addresses the aforementioned limitation of Study 1 by directly investigating the relationship between messages' moral framing and audiences' moral values.

Study 2a

In Study 2a, we developed a set of stimuli consisting of social media posts about either true or false news headlines. These posts were framed to either aligned with Individualizing or Binding values or were framed in nonmoral terms.

Method

Participants. We recruited 804 U.S. American Twitter users from the Prolific subject pool who, according to their responses to the Prolific prescreening questionnaire, were U.S. residents, used Twitter at least once a month, and had posted on Twitter at least 1–3 times in the last 12 months. Our sample was stratified by gender ($\frac{1}{2}$ female, $\frac{1}{2}$ male) and political orientation ($\frac{1}{3}$ liberal, $\frac{1}{3}$ moderate, $\frac{1}{3}$ conservative). We excluded participants who failed at least one of three attention checks or whose responses conflicted with their responses to the Prolific prescreening questionnaire. This left a final sample of 615 participants ($Mdn = 32$ years, age range: 18–79 years; 304 women, 305 men, 6 other) of whom 205 identified as conservative, 205 identified as moderate, and 205 identified as

liberal. As Figure 1 shows, our sample spanned the whole spectrum of political orientation.

Stimuli. To create the stimuli set, we selected 51 news headlines (23 true, 28 false) from the fact-checking website snopes.com and created three social media posts for each news headline. Social media posts were designed to look like Twitter posts, with information unrelated to the study (e.g., the date, the poster's identity, and profile picture) blurred. Specifically, we used *moral reframing* (Feinberg & Willer, 2019) to create, for each headline, three posts that commented on the headline: one post that appealed to Binding values (Loyalty, Authority, Purity), one post that appealed to Individualizing values (Care, Equality), and one post that avoided moral sentiment. For each headline, we created posts that all either expressed negative sentiment (27) or positive sentiment (24). This resulted in 51 (news headline) $\times$ 3 (moral framing) = 153 social media posts.

For example, we created three social media posts for the true news headline: "Portland Named a New Bridge After 'The Simpsons' Ned Flanders" (MacGuill, 2021). Two posts commented on the headline in a way that appealed either to Binding values (e.g., "I read this article and I can't believe it! This bridge should be named after a great American patriot, not a cartoon character!") or to Individualizing values ("I read this article and can't believe it. We have so many civil rights leaders who go nameless and we give it to another white man!"). Another post commented on the headline in nonmoral terms (e.g., "I read this article and am surprised—a bridge named after a Simpsons character?! Ridiculous! People have too much time on their hands!"). For this headline, all posts expressed negative sentiment.

Procedure. After agreeing to participate, participants responded to three questions that mirrored the questions in the Prolific prescreening questionnaire. Provided that participants' answers matched their pre-screening questionnaire, they were informed that the following pages would showcase social media posts, each containing a news headline and a user's written commentary. We informed them that some details about the posts, such as who posted it and when, were omitted. Participants were instructed to

answer each question as if they had come across the post while using social media (e.g., Twitter or Facebook).

Participants then responded to randomly sampled social media posts, none of which were about the same news headline. For each post, participants answered several questions about the shared headline, and the post about the shared headline. They also rated how likely they would be to share the post if they came across it. We used a planned missingness design so that each participant responded to 6 of 153 posts and each post was rated by 15–35 participants. After responding to six posts, participants completed the MFQ-2 and the demographic measures. On the final page, participants read that they had seen both real and fake news headlines and were provided with a table of all headlines, showing which ones were true and false.

Measures. For each social media post, participants rated how much the post aligned with their values on a 5-point Likert scale (1 = *strongly opposed to my values*, 5 = *strongly aligned with my values*). Then, participants completed the 36-item moral foundations questionnaire (MFQ-2, Atari et al., 2022) which assesses to what extent participants endorse moral concerns about Care (e.g., "We should all care for people who are in emotional pain."), Equality (e.g., "The world would be a better place if everyone made the same amount of money."), Proportionality (e.g., "I think people who are more hard-working should end up with more money."), Loyalty (e.g., "It upsets me when people have no loyalty to their country."), Authority (e.g., "I believe that one of the most important values to teach children is to have respect for authority."), and Purity (e.g., "I believe chastity is an important virtue."; 1 = *does not describe me at all*, 5 = *describes me extremely well*). Items were presented in random order with three additional attention checks embedded within the questionnaire (e.g., "To show that you are paying attention and giving your best effort, please select 'moderately describes me'."). In addition to the aforementioned two measures that were central to the purpose of Study 2a, participants in Study 2a also completed a subset of additional measures used in Study 2b to facilitate

exploratory analysis and piloting (see an overview of our measures in section B of the supplemental materials).

Results

To select stimuli for Study 2b, we correlated participants' responses to the question, "How much does the post the user has written about the headline align with your values?", with their endorsement of Binding and Individualizing values. To that end, we calculated an index of Binding values by averaging a participants' endorsement of the Loyalty, Authority, and Purity foundations and an index of Individualizing values by averaging a participants' endorsement of the Care and Equality foundations. We selected those news headlines for which (1) for posts framed with Binding values, the correlations of participants' ratings of the extent that the post aligned with their moral values were maximally more positive for Binding compared to Individualizing values, (2) for posts framed with Individualizing values, the correlations of participants' moral alignment ratings were maximally more positive for Individualizing compared to Binding values, and (3) for posts with nonmoral framing, the correlations of participants' moral alignment ratings with participants' Binding and Individualizing values were smallest. Using this criterion, we selected the 5 x 2 (positive/negative sentiment) x 2 (true/false headline) = 20 best sets of 3 stimuli (Binding/Individualizing/nonmoral framing) for use in future studies (see Table A2 in the supplemental materials). In this way, Study 2a resulted in a paradigm that facilitates the investigation of how moral framing affects responses to shared social media content.

Study 2b

In Study 2b, we used the newly developed paradigm to test hypotheses about the relationships between moral framing, moral values, and responses to shared social media content. We tested two preregistered hypotheses, predicting that respondents would be more likely to share a social media post about a news headline if the framing of the post aligned with their moral values (Hypothesis 1) and that they would do so *because* they

agreed with the post and *because* it aligned with their moral values (Hypothesis 2).

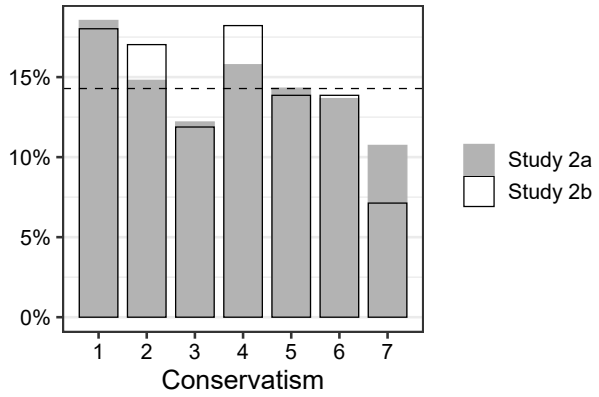
Method

We preregistered the sample size as well as all hypotheses, inclusion/exclusion criteria, statistical models, measures, and manipulations (https://osf.io/f7r8d/?view_only=7e4b1b5e3c574be6848664235fbd41ca). We made all materials, data, and analysis scripts available online (https://osf.io/z25tc/?view_only=0141845d12024a2cbdbd0f71f77f23a8).

Participants. We recruited 641 U.S. American Twitter users from the Prolific subject pool who, according to their responses to the Prolific prescreening questionnaire, were U.S. residents, used Twitter at least once a month, who had posted on Twitter at least 1–3 times in the last 12 months, and who had not participated in Study 2a. We excluded 136 participants who failed at least one of three attention checks or whose responses in our survey conflicted with their responses to the Prolific prescreening questionnaire. We had preregistered that we would recruit a sample of 540 eligible participants, stratified by gender ($\frac{1}{2}$ female, $\frac{1}{2}$ male) and self-identified political orientation ($\frac{1}{3}$ liberal, $\frac{1}{3}$ moderate, $\frac{1}{3}$ conservative). We found, however, that, after recruiting 145 conservative participants, we exhausted the pool of eligible conservative participants in the Prolific subject pool and concluded data collection. This left a final sample of 505 participants ($Mdn = 32$ years, age range: 18–79 years; 231 women, 269 men, 5 other) of whom 145 identified as conservative, 180 identified as moderate, and 180 identified as liberal. As Figure 1 shows, our sample spanned the whole spectrum of political orientation.

Procedure. We used a planned missingness design that allowed both within-subject and between-subjects comparisons. In total, we included 2 (headline: true, false) $\times$ 2 (post: positive, negative sentiment) $\times$ 5 = 20 news headlines selected in Study 2a (Table A2). In total, we included 3 (Binding, Individualizing, nonmoral framing) $\times$ 20 (news headlines) = 60 social media posts. Each participant responded to six randomly sampled social media posts, none of which were based on the same news headline. That is,

Figure 1
Distribution of political orientation across samples

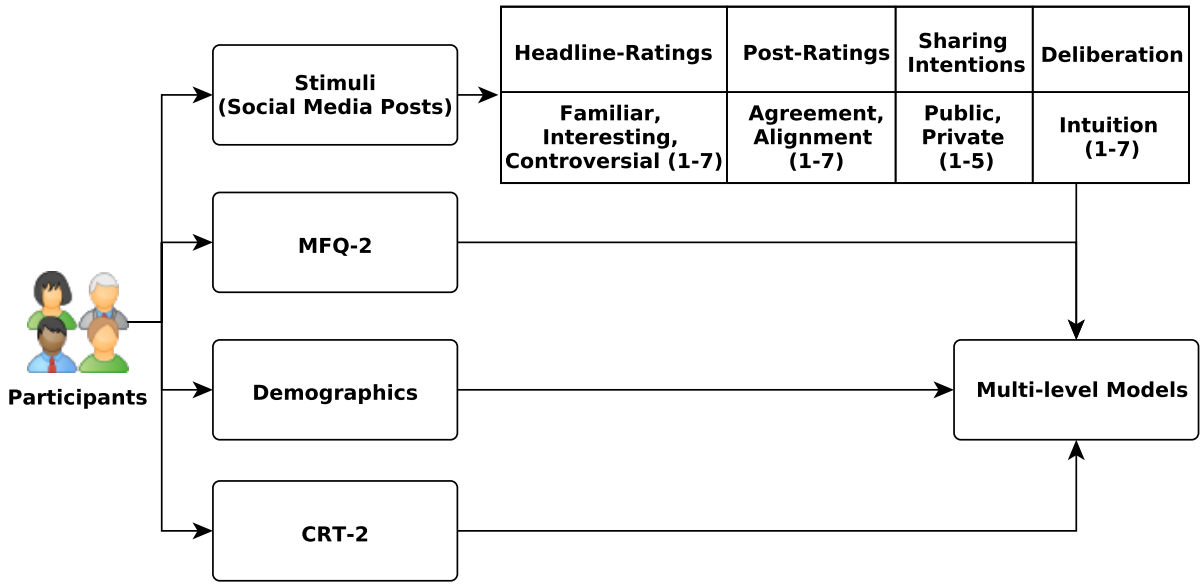


Note. "How would you describe your political beliefs?" (1 = liberal, 7 = conservative). Dashed line shows proportions expected under a uniform distribution.

the same participant responded to posts using Binding, Individualizing, or nonmoral framings (within-subject comparison) but different participants respond to posts using different framings of the same headline (between-subject comparison). Each post was rated by 33–66 participants. See Figure 2 for an illustration of the general study procedures and see Figure 3 for an example of how the stimuli from Table A2 were presented to the participants. A summary of the stimuli presentation as well as the survey items for the post and headline ratings can be found in section A and B of the supplemental materials.

Figure 2

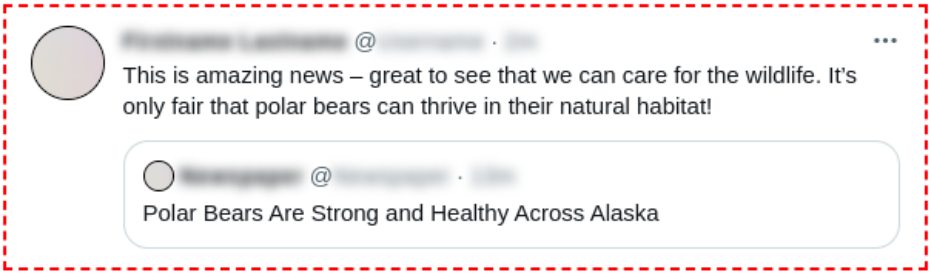
Illustration of Study flow



Note. Participants are presented with stimuli (social media post with a news headline and text about news headline). Participants then give ratings on the headline-level, post-level, indicate their sharing intentions, and deliberation over sharing (Study 3). After stimulus presentation, participants complete the MFQ-2, CRT-2 (Study 3), and demographic questions.

Figure 3

Exemplary stimulus presentation (shared news headline).



Note. After stimulus presentation, participants are asked to give ratings on the headline-level (e.g., believability), post-level (e.g., surprising), and indicate their sharing intentions.

Study 2b followed the same procedure as Study 2a, except that, after responding to six posts, participants were again shown each of the posts and asked to indicate which of the headlines they thought were true or false.

Measures. For each social media post, we used bipolar adjective ratings to measure how unbelievable–believable, uncontroversial–controversial, unsurprising–surprising, uninteresting–interesting, and negative–positive a participant rated the news headline as well as the post about the news headline (1–7). We also measured how much a participant agreed or disagreed with the post that was written about the headline (1 = *strongly disagree*, 5 = *strongly agree*) and how much the post that was written about the headline aligned with the participant's values (1 = *strongly opposed to my values*, 5 = *strongly aligned with my values*).

For each social media post, we also recorded how likely participants would be to share the post publicly on their social media feed; 'like' the post; share the post in a private message, text message, or email; and talk about the post or headline in an offline conversation (1 = *very unlikely*, 5 = *very likely*). We calculated an index of sharing intentions by averaging each participants' responses to the four items for each post they responded to ($\alpha = 0.86$). Participants were also asked to indicate whether they believed each headline to be true or false (1 = *true*, 0 = *false*).

In addition, participants completed the 36-item MFQ-2 (Atari et al., 2022) to measure how much they endorsed Binding (Loyalty, Authority, Purity) and Individualizing (Care, Equality) values. Participants also responded to demographic questions about their gender, education, and political beliefs.

Analysis Strategy. We ran a series of multilevel linear regression models that estimated participants' z-standardized sharing intentions as a function of various predictor variables. Model 0 did not include any predictor variables and estimated sharing intentions as a function of a fixed intercept and three varying (random) intercepts that accounted for variance across posts, headlines, and participants. Model 1 extended Model 0 by estimating sharing intentions as a function of ratings of how believable, controversial, surprising, interesting, and positive a headline was perceived to be. We modeled headline-level predictor variables with the fixed effect of the z-standardized average ratings

of each headline and with the fixed and varying (across headlines) effect of each participant's z-standardized deviation from the average rating for each headline. Model 2 extended Model 0 by estimating sharing intentions as a function of ratings of how controversial, surprising, interesting, and positive a post about a headline was perceived to be. We modeled post-level predictor variables with the fixed effect of the z-standardized average ratings of each post and with the fixed and varying (across posts) effect of each participant's z-standardized deviation from the average rating for each post. Model 3 mirrored Model 2 but included only post-level ratings of how much participants agreed with the post, how much the post aligned with their values, and the interaction between the two. Model 4 extended Model 0 by estimating sharing intentions as a function of participants' endorsement of Binding, Individualizing, and proportionality values. We modeled participant-level predictor variables with the fixed effect and varying (across headlines) effect of the participants' z-standardized moral concerns, the dummy-coded framing of each post, and the interaction between the two. Model 5 mirrored Model 4 but included participants' z-standardized conservatism instead of their endorsement of moral concerns. Lastly, we ran a multivariate multilevel linear regression model (mediation) to estimate indirect effects of moral concerns on sharing intentions via ratings of how much participants agreed with each post and how much it aligned with their moral values. We estimated and evaluated these models analogously to Study 1, using the 'brms' R package and 10-fold cross-validated ELPD scores.

Results

Preregistered Analyses. Table 3 compares each model's out-of-sample prediction accuracy to that of the null model without predictors (M0) and that of the other models with predictors (M1–M5). Supporting Hypothesis 1, Model 4—that included participants' endorsement of Binding and Individualizing values and their interactions with the moral framing of each social media post as predictor variables—predicted sharing

Table 3
Comparison of preregistered models estimating sharing intentions as a function of various predictor variables

			z					
Model	Description	R^2	M0	M1	M2	M3	M4	M5
M0	No Predictors	.00	-	-13.11	-15.82	-11.08	-3.53	-1.16
M1	Headline-Level Ratings	.15	13.11	-	-4.16	-0.20	8.83	11.73
M2	Post-Level Ratings	.21	15.82	4.16	-	3.47	12.41	15.36
M3	Agreement/Alignment	.18	11.08	0.20	-3.47	-	8.30	11.00
M4	Moral Concerns	.08	3.53	-8.83	-12.41	-8.30	-	2.89
M5	Political Orientation	.02	1.16	-11.73	-15.36	-11.00	-2.89	-

Note. R^2 *is a Bayesian analogue to the proportion of within-sample variance explained by a model (not considering varying effects).* z *is the difference in out-of-sample prediction accuracy between two models divided by its standard error* ($z = \Delta_{ELPD}/SE$).

intentions more accurately than Model 0 ($\Delta_{ELPD} = 59.11, SE = 16.73, z = 3.53$). As hypothesized, participants' endorsement of Binding values predicted greater sharing intentions in the Binding framing condition ($\beta = 0.26, [0.16, 0.36]$) than in the Individualizing framing condition ($\beta = 0.14, [0.03, 0.24]; \Delta\beta = 0.12, [0.03, 0.21]$) and, to a lesser extent, in the nonmoral framing condition ($\beta = 0.20, [0.10, 0.30]; \Delta\beta = 0.06, [-0.03, 0.15]$). In other words, participants with Binding values had greater sharing intentions for posts framed with Binding values (aligned) than posts with Individualizing values (misaligned).

Likewise, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition ($\beta = 0.23, [0.16, 0.31]$) than in the Binding framing condition ($\beta = 0.07, [-0.01, 0.14]; \Delta\beta = 0.16, [0.09, 0.24]$) and, to a lesser extent, in the nonmoral framing condition ($\beta = 0.14, [0.06, 0.21]; \Delta\beta = 0.10, [0.01, 0.18]$). In other words, participants with individualizing values had greater sharing intentions for posts framed with individualizing values (aligned) than nonmoral posts (neutral) and posts with Binding values (misaligned). Participants' endorsement of proportionality concerns was unrelated to sharing intentions

in all three framing conditions
($\beta = 0.00, [-0.09, 0.09]; \beta = -0.05, [-0.14, 0.04]; \beta = -0.03, [-0.12, 0.06]$). [5]

Notably, Model 4 predicted sharing intentions more accurately than Model 5 ($\Delta_{ELPD} = 50.90, SE = 16.90, z = 3.01$) which predicted sharing intentions as a function of political orientation instead of moral concerns. Taken together, these findings emphasize the facilitatory effect of targeting people's moral values on sharing (mis)information. Models that estimated sharing intentions as a function of headline-level ratings (M1; $z = 8.96$), of post-level ratings (M2; $z = 12.52$), or of post-level alignment and agreement ratings (M3; $z = 8.39$) made more accurate out-of-sample predictions than the model that estimated sharing intentions as a function of moral concerns and their interaction with moral framing (M4). Across the three models (M1–M3), the most important predictors were to what extent a participant rated the headline to be interesting (M1: $\beta = 0.27, [0.22, 0.32]$) and believable (M1: $\beta = 0.11, [0.06, 0.15]$); rated the post to be interesting (M2: $\beta = 0.34, [0.31, 0.38]$) and positive (M2: $\beta = 0.13, [0.09, .16]$); and agreed with the post (M3: $\beta = 0.16, [0.11, 0.21]$), considered the post to align with their moral values (M3: $\beta = 0.22, [0.17, 0.28]$), or both (M3: $\beta = 0.14, [0.11, 0.16]$). These findings were, perhaps, not surprising as the predictor variables included in those models, especially Model 2, were more proximal to our outcome variable and related to core motives of using social media (i.e., eliciting social connection/interactions: Al-Saggaf and Nielsen (2014), Sung et al. (2016), and Wu and Atkin (2017)). Nevertheless, our findings show that both perceived alignment of shared content and participant values, as well as the specific match between moral framing and moral values, have a significant facilitating effect on sharing intentions.

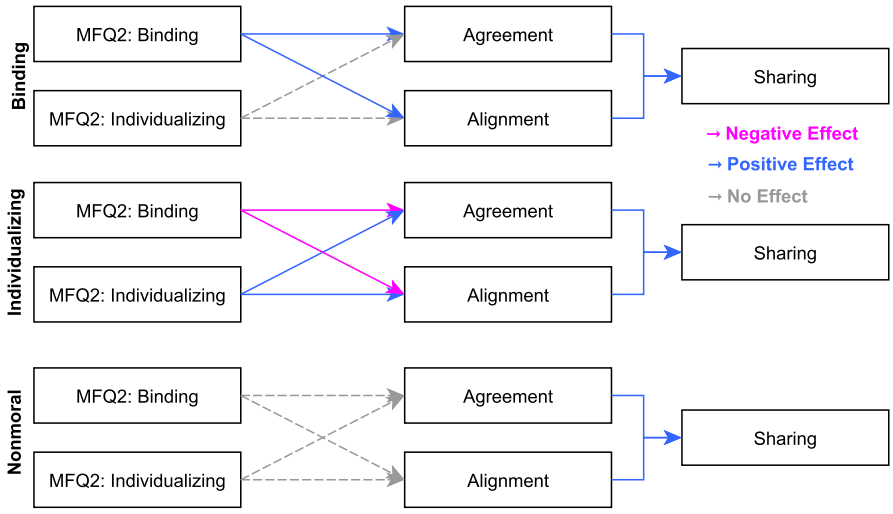
To test Hypothesis 2, we estimated a Bayesian multilevel mediation model and compared, across the three moral framing conditions, the total indirect effects of

[5] Importantly, all effects held when controlling for headline veracity, which did not have a significant effect ($\beta = -0.03, [-0.13, 0.07]$) on sharing intentions. See Table F3 for effect sizes when controlling model 4 for headline veracity.

participants' endorsement of Binding and Individualizing values on sharing intentions via their ratings of how much they agreed with the post, how much the post aligned with their moral values, and the interaction of the two ratings, while controlling for headline veracity. Figure 4 provides an overview of the observed relationships. Supporting Hypothesis 2, we found that, first, participants' endorsement of Binding values had a positive indirect effect on sharing intentions in the Binding framing condition ($\beta = .31, [.22, .41]$) but a negative indirect effect in the Individualizing framing condition ($\beta = -.12, [-.20, -.04]$); second, participants' endorsement of Individualizing values had a positive indirect effect on sharing intentions in the Individualizing framing condition ($\beta = .30, [.21, .38]$) but no indirect effect in the Binding framing condition ($\beta = -.03, [-.10, .04]$); and third, participants' endorsement of Binding ($\beta = .05, [-.03, .13]$) and Individualizing ($\beta = .03, [-.06, .11]$) values had no indirect effect in the nonmoral framing condition.

Figure 4

Results from the preregistered mediation analysis



Note. Results show that there is a positive mediation (blue color) for match of moral framing and moral values, and no effect (grey) or a negative effect (red) for a mismatch.

Therefore, supporting the original hypothesis of Study 1, the results of Study 2b showed that an alignment of moral framing and moral values (Binding values and Binding framing or Individualizing values and Individualizing framing) indeed increases sharing of

social media posts, even when controlled for veracity. In other words, aligning a post's framing with a user's core values will increase sharing intentions independent of the post containing true or false content. Importantly, we also found that a match of framing and values predicts sharing intentions more accurately than other related variables, such as political ideology. Additionally, our findings that moral framing and political ideology interact to predict sharing (Study 1) but that moral values predict sharing more accurately than political ideology (Study 2) suggests that political ideology and moral values are linked but distinct concepts.

Study 3

The results of Studies 1 and 2 support the hypothesis that aligning a social media post's moral framing with a user's core values increases sharing intentions but leave open the underlying mechanism. For instance, matching moral values and message framings could elicit a moral-emotional response that facilitates message sharing by distracting participants from deliberating and thus from carefully judging whether the post is false or implausible. If so, then the effect of aligning moral values and message framing should be mediated by deliberation. Alternatively, participants could be motivated by their intuitions of right and wrong that accompany moralized posts and supersede accuracy concerns. In this case, there should not be an effect of deliberation on sharing intentions. To test these hypotheses, Study 3 replicates Study 2b in a pre-registered experiment and includes measures of deliberation.

We first replicate Study 2b and its original hypotheses, predicting that respondents would be more likely to share a social media post about a news headline if the framing of the post aligns with their moral values (Hypothesis 1). We then investigate whether these findings can be explained by deliberation; if the effect of aligning posts' moral framing and respondents' moral values is mediated by how much they deliberate about sharing the post (Hypothesis 2) and whether susceptibility to this effect is moderated by trait-level

analytical thinking (Hypothesis 3). As done in previous works, we utilize the Cognitive Reflection Test (CRT-2; Thomson and Oppenheimer (2016)) as a trait-level measure of analytical thinking. We also directly measure deliberation over sharing a post via ratings of how much a participant's decision is guided by deliberation or intuition.

We preregistered the sample size as well as all hypotheses, inclusion/exclusion criteria, statistical models, measures, and manipulations[6] and made all materials, data, and analysis scripts available online[7].

Method

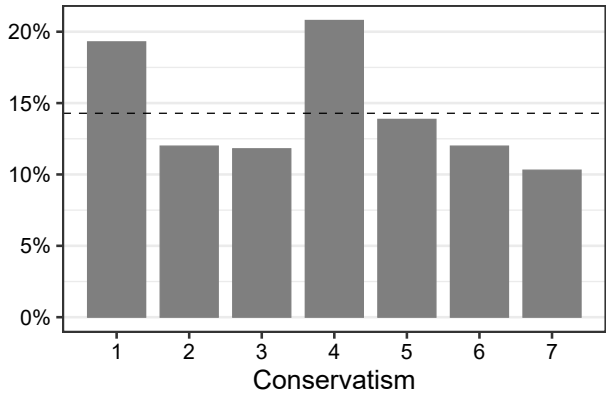
Participants. We recruited 676 U.S. American Twitter users from the Prolific subject pool who, according to their responses to the Prolific prescreening questionnaire, were U.S. residents, used Twitter at least once a month, posted on Twitter at least 1–3 times in the last 12 months, and who had not participated in Study 2a or 2b. We excluded participants who failed at least one of three attention checks or whose responses conflicted with their responses to the Prolific prescreening questionnaire. We had preregistered that we would recruit a sample of 540 eligible participants, stratified by gender ($\frac{1}{2}$ female, $\frac{1}{2}$ male) and self-identified political orientation ($\frac{1}{3}$ liberal, $\frac{1}{3}$ moderate, $\frac{1}{3}$ conservative). After excluding participants with failed attention checks or missing data, we were left with a final sample of 533 participants ($Mdn = 32$ years, age range: 18–75 years; 265 women, 256 men, 12 other) of whom 178 identified as conservative, 177 identified as moderate, and 178 identified as liberal. As Figure 5 shows, our sample spanned the spectrum of political orientation.

Procedure. We used the same planned missingness design from Study 2b that allowed both within-subject and between-subjects comparisons. We included the same 2 (headline: true, false) $\times$ 2 (post: positive, negative sentiment) $\times$ 5 = 20 news headlines

[6] Preregistration Link

[7] OSF Repository Link

Figure 5
Distribution of political orientation across samples



Note. "How would you describe your political beliefs?" (1 = liberal, 7 = conservative). Dashed line shows proportions expected under a uniform distribution.

selected in Study 2a (Table A2) and used in Study 2b. In total, we included 3 (Binding, Individualizing, nonmoral framing) × 20 (news headlines) = 60 social media posts. Each participant responded to six randomly sampled social media posts, none of which were based on the same news headline. That is, the same participant responded to posts using Binding, Individualizing, or nonmoral framings (within-subject comparison) but different participants responded to posts using different framings of the same headline (between-subject comparison). Each post was rated by 33–66 participants.

Study 3 followed the same procedure as Study 2b, except we removed the items asking participants whether the presented headlines were true or false because these measures were only used for exploratory analysis outside of the scope of this work.

Measures. We collected the same post and headline-level ratings as in Study 2b, such as how believable, controversial, surprising, interesting, and negative participants rated the social media posts and news headlines, how much they agreed with a post, how much it aligned with their values, and their sharing intentions. To increase robustness of our estimates, we added a measure of headline familiarity (unfamiliar – familiar; 1–7) as an additional control variable for our analyses because familiarity is linked to perceived accuracy of news due to fluency effects (Pennycook & Rand, 2020; Schwarz et al., 2016;

Swire et al., 2017). Participants also indicated to what extent they deliberated or used intuition when deciding to share or not to share a post (bipolar items; intuition – deliberation, $\alpha = 0.65$).

Participants again completed the 36-item MFQ-2 and responded to the same demographic questions about their gender, education, and their political beliefs. Lastly, participants completed the Cognitive Reflection Task 2 (Thomson & Oppenheimer, 2016), which measures to what extent participants generally think analytically or intuitively.

Analysis Strategy. Similar to Study 2b, we ran a series of multilevel linear regression models that estimated participants' z-standardized sharing intentions as a function of various predictor variables.

First, we replicated the models from Study 2b: Model 0 did not include any predictor variables and estimated sharing intentions as a function of a fixed intercept and three varying (random) intercepts that accounted for variance across posts, headlines, and participants. Model 1 extended Model 0 by estimating sharing intentions as a function of ratings of how believable, controversial, surprising, interesting, familiar and positive a headline was perceived to be. We modeled headline-level predictor variables with the fixed effect of the z-standardized average ratings of each headline and with the fixed and varying (across headlines) effect of each participant's z-standardized deviation from the average rating for each headline. Model 2 extended Model 0 by estimating sharing intentions as a function of ratings of how controversial, surprising, interesting, and positive a post about a headline was perceived to be. We modeled post-level predictor variables with the fixed effect of the z-standardized average ratings of each post and with the fixed and varying (across posts) effect of each participant's z-standardized deviation from the average rating for each post. Model 3 mirrored Model 2 but included only post-level ratings of how much participants agreed with the post, how much the post aligned with their values, and the interaction between the two. Model 4 extended Model 0 by estimating sharing intentions as a function of participants' endorsement of Binding, Individualizing, and proportionality

values as well as their interactions with posts' moral framing. We modeled participant-level predictor variables with the fixed effect and varying (across headlines) effect of the participants' z-standardized moral concerns, the dummy-coded framing of each post, and the interaction between the two. Model 5 mirrored Model 4 but included participants' z-standardized conservatism instead of their endorsement of moral concerns. Lastly, we ran a multivariate multilevel linear regression model (mediation) to estimate the indirect effects of moral concerns on sharing intentions via ratings of how much participants agreed with each post and how much it aligned with their moral values.

Second, we ran a multivariate multilevel linear regression model (mediation) to estimate the indirect effects of moral concerns on sharing intentions via ratings of how much participants deliberated to share each post. We also included analytical thinking (CRT-2) in this model as a potential moderator (higher trait-level analytical thinking should reduce susceptibility to the effect of moral framing).

Analogous to Study 1 and Study 2b, we used the 'brms' R package to estimate the generalized linear multilevel models and used 10-fold cross-validated ELPD scores for model comparison.

Results

Preregistered Analyses. Replicating the analysis in Study 2b (Hypothesis 1), Table 4 compares each model's out-of-sample prediction accuracy to that of the null model without predictors (M0) and that of the other models with predictors (M1–M5). Supporting Hypothesis 1, Model 4—that included participants' endorsement of Binding and Individualizing values and their interactions with the moral framing of each social media post as predictor variables—predicted sharing intentions more accurately than Model 0 ($\Delta_{ELPD} = 59.66, SE = 16.91, z = 3.68$). As hypothesized, participants' endorsement of Binding values predicted greater sharing intentions in the Binding framing condition ($\beta = 0.26, [0.17, 0.34]$) than in the Individualizing framing condition

Table 4
Comparison of preregistered models estimating sharing intentions as a function of various predictor variables

			z					
Model	Description	R^2	M0	M1	M2	M3	M4	M5
M0	No Predictors	.00	-	-15.28	-14.93	-13.41	-3.68	0.06
M1	Headline-Level Ratings	.16	15.28	-	-0.70	-1.48	10.94	14.40
M2	Post-Level Ratings	.18	14.93	0.70	-	-0.92	11.60	14.57
M3	Agreement/Alignment	.22	13.41	1.48	0.92	-	11.10	13.67
M4	Moral Concerns	.07	3.68	-10.94	-11.60	-11.10	-	3.66
M5	Political Orientation	.02	-0.06	-14.40	-14.57	-13.67	-3.66	-

Note. R^2 *is a Bayesian analogue to the proportion of within-sample variance explained by a model (not considering varying effects).* z *is the difference in out-of-sample prediction accuracy between two models divided by its standard error* ($z = \Delta_{ELPD}/SE$).

($\beta = 0.11, [0.02, 0.20]; \Delta\beta = 0.14, [0.05, 0.23]$) and, to a lesser extent, in the nonmoral framing condition ($\beta = 0.15, [0.06, 0.24]; \Delta\beta = 0.11, [0.02, 0.15]$). In other words, participants with Binding values had greater sharing intentions for posts framed with Binding values (aligned) than posts with Individualizing values (misaligned) or without moral framing (neutral).

Likewise, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition ($\beta = 0.26, [0.18, 0.34]$) than in the Binding framing condition ($\beta = 0.11, [0.03, 0.19]; \Delta\beta = 0.15, [0.08, 0.22]$) and, to a lesser extent, in the nonmoral framing condition ($\beta = 0.13, [0.06, 0.21]; \Delta\beta = 0.13, [0.05, 0.20]$). In other words, participants with Individualizing values had greater sharing intentions for posts framed with Individualizing values (aligned) than nonmoral posts (neutral) and posts with Binding values (misaligned). Participants' endorsement of proportionality concerns was unrelated to sharing intentions in all three framing conditions ($\beta = 0.01, [-0.08, 0.10]; \beta = -0.01, [-0.10, 0.08]; \beta = 0.02, [-0.07, 0.11]$). Importantly, all effects held when controlling for headline veracity, which did not have a significant effect ($\beta = -0.03, [-0.13, 0.07]$) on sharing intentions, and headline familiarity, which had a

positive effect on sharing intentions ($\beta = 0.18, [0.14, 0.21]$). See Table 5 for a comparison of effect sizes for model 4 with and without controls for headline veracity and familiarity.

Table 5
Effect sizes for Model 4 when controlling for headline veracity and familiarity

Values – Framing Condition	β[CI]	$\Delta\beta_{no\ control}$
Binding – Binding	.25 [.17, .33]	-0.01
Binding – Individualizing	.09 [.01, .18]	-0.02
Binding – Nonmoral	.14 [.05, .23]	0.01
Individualizing – Binding	.09 [.02, .16]	-0.02
Individualizing – Individualizing	.24 [.16, .32]	-0.02
Individualizing – Nonmoral	.12 [.05, .20]	-0.01
Proportionality – Binding	.01 [-.08, .10]	0.00
Proportionality – Individualizing	.00 [-.09, .09]	0.01
Proportionality – Nonmoral	.03 [-.06, .11]	0.01
$\Delta\beta_{Binding:Binding-Individualizing}$	.16 [.07, .24]	0.02
$\Delta\beta_{Binding:Binding-Nonmoral}$	.11 [.02, .19]	0.00
$\Delta\beta_{Individualizing:Individualizing-Binding}$	.15 [.08, .22]	0.00
$\Delta\beta_{Individualizing:Individualizing-Nonmoral}$	.12 [.04, .19]	-0.01

Note. Table shows the effect sizes for a given moral value in a given framing condition, as well as the difference in effect sizes for a moral value across framing conditions (last 4 rows). Table shows that the reported effect sizes in Study 3 hold when controlling for headline veracity.

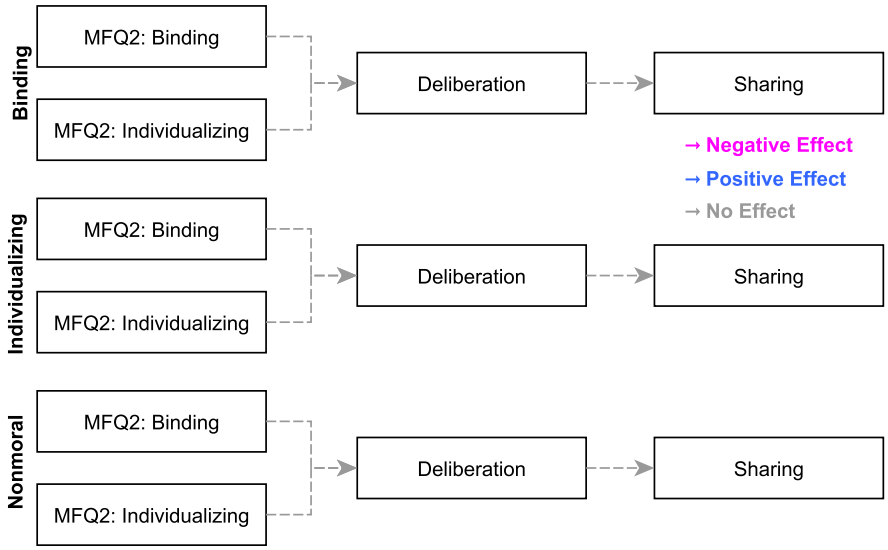
Consistent with Study 2b, Model 4 predicted sharing intentions more accurately than Model 5 ($\Delta_{ELPD} = 60.22, SE = 16.46, z = 3.66$) which predicted sharing intentions as a function of political orientation instead of moral concerns. Overall, Study 3 successfully replicated the facilitatory effect of targeting people's moral values on (mis)information sharing. Consistent with Study 2b, models that estimated sharing intentions as a function of headline-level ratings (M1; $z = 10.94$), of post-level ratings (M2; $z = 11.60$), or of post-level alignment and agreement ratings (M3; $z = 11.10$) made more accurate out-of-sample predictions than the model that estimated sharing intentions as a function of moral concerns and their interaction with moral framing (M4). Across the three models (M1–M3), the most important predictors were, consistent across both studies, to what extent a participant rated the headline to be interesting (M1: $\beta = 0.26, [0.21, 0.32]$), believable (M1: $\beta = 0.13, [0.09, 0.17]$), familiar (additional in Study 3; M1:

$\beta = 0.10, [0.07, 0.13]$); rated the post to be interesting (M2: $\beta = 0.35, [0.30, 0.39]$) and positive (M2: $\beta = 0.10, [0.07, 0.13]$); and agreed with the post (M3: $\beta = 0.17, [0.12, 0.22]$), considered the post to align with their moral values (M3: $\beta = 0.28, [0.23, 0.33]$), or both (M3: $\beta = 0.12, [0.09, 0.14]$).

To test Hypothesis 2, which is an alternative explanation of our findings in Study 2b – that participants' sharing behavior is not motivated by the desire to align with their own values but that they are instead distracted from estimating message veracity or plausibility – we investigated whether deliberation about sharing a post mediated the effect of aligning a post's moral framing and participants' moral values. To that end, we estimated a Bayesian multilevel mediation model with ratings of how much participants deliberated about sharing a post would mediate the effect of moral framing and moral values on sharing intentions, while controlling for headline veracity and familiarity. Figure 6 provides an overview of the observed relationships. We found no evidence for a mediation. Alignment of moral values and moral framing did not predict less deliberation ($\beta = 0.02, [-0.02, 0.07]; \beta = -0.00, [-0.05, 0.04]$) and, importantly, deliberation did not predict lower sharing intentions for false news compared to true news ($\beta = 0.02, [-0.03, 0.08]$). Furthermore, analytical thinking did not moderate the effect of aligning moral values and moral framing ($\beta = -0.01, [-0.06, 0.04]; \beta = -0.03, [-0.08, 0.03]$), meaning that analytical thinking did not reduce susceptibility to moral framing (Hypothesis 3). We also ran an identical mediation model with response time for sharing a post as an alternative deliberation measure with shorter response time indicating faster, more intuitive thinking. We again found no mediation, with no effect of moral framing and values on response time ($\beta = -0.04, [-0.13, 0.05]; \beta = 0.02, [-0.03, 0.07]$) and longer response time (indicating deliberation) did not predict lower sharing intentions of false news compared to true news ($\beta = 0.01, [-0.03, 0.05]$). For a more detailed analysis of analytical thinking see section D in the supplemental materials.

Figure 6

Results from the preregistered mediation analysis of the effect of aligning moral framing and moral values via deliberation



Note. Figure shows that deliberation does not mediate the effect of moral alignment on sharing intentions. Importantly, there is no effect of deliberation on sharing intentions.

Summary. Supporting the original hypotheses of Study 1 and Study 2b, Study 3 confirmed that an alignment of a post's moral framing to users' moral values indeed increased sharing of social media posts, even when controlling for veracity and familiarity. In other words, aligning a post's framing with a user's moral values will increase sharing intentions independent of the post containing true or false content, and how familiar the content is. We also consistently found that a match between a user's moral values and posts' moral framing predicted sharing intentions more accurately than other related variables, such as political ideology. Furthermore, our results showed that matching post framing and user values increases sharing intentions independent of deliberative thinking. Specifically, alignment of posts' framing and users' moral value did not reduce deliberation over sharing a post and deliberation did not impact the intention to share misinformation. Finally, trait-level analytical thinking did not moderate the effect of moral alignment, did not affect sharing of misinformation and did not increase plausibility concerns, except for

nonmoral posts (see additional analyses in supplemental materials).

General Discussion

We investigated the role of moral framing, specifically the match or mismatch of an individual's moral values and a message's moral framing, on belief and sharing of social media posts. Across three studies, one large-scale analysis of real-world conversations on Twitter and two behavioral experiments, we found that a match of framing and values leads to increased sharing of (mis)information. Crucially, these effects were found while controlling for message veracity and message familiarity.

These results are relevant for the current debate on psychological drivers of misinformation spread. Our findings indicate that it is not just moral content but rather *matched* moral content that matters. Importantly, our experimental manipulation (i.e., framing a message by using language centered around Individualizing or Binding values) was independent of the message's core contents, such as its main arguments or partisanship. For example, a headline about the State Department charging Americans for evacuation flights could be framed using Individualizing Language (e.g.,"It is unfair that only the rich get saved") or Binding Language (e.g.,"The government is betraying its poor citizens") without changing the main argument that the government shouldn't charge for evacuations, the negative sentiment, or left-leaning partisanship. We created our stimuli through careful crafting of matched messages while staying away from obviously partisan content and counterbalancing moral content across headline veracity. Since we avoided confounding message content and moral framing with political ideology, the absence of an effect of political ideology must not be misinterpreted as partisanship in messages or individual differences in conservatism not playing a role in (mis)information sharing. Rather, our results suggest that the effects of such variables are driven by underlying moral values. In the real world, where partisan messages are frequently accompanied by moralized language and arguments, and where partisan groups differ in their endorsement

of moral values (Graham et al., 2009; Kivikangas et al., 2021), this can indeed lead to partisan differences in misinformation sharing as observed by prior research (Kaplan et al., 2021; Van Bavel & Pereira, 2018; Winkielman et al., 2012).

Past work on misinformation has shown that more deliberative, analytical reasoning often leads to less sharing of misinformation, indicating that more analytical individuals might be able override initial sharing intentions (Bronstein et al., 2019; Pennycook & Rand, 2019a, 2019b, 2019c). However, in Study 3, analytical and "lazy" thinkers did not differ in their sharing of misinformation and how much they relied on plausibility cues. Furthermore, deliberation over sharing a post did not predict lower sharing intentions of misinformation and did not mediate the effect of aligning moral framing and moral values on misinformation sharing intentions, meaning that moral framing did not simply distract participants from accuracy cues. It is possible that the effect of analytical thinking is restricted to contexts that do not strongly evoke values, group identities, and threats thereof (e.g., see the following work for failures to replicate the effect of analytical thinking or failures to exclude motivational drivers for highly politically polarized, moralized, and identity-relevant issues: Lee et al. (2020), Osmundsen et al. (2021), Pretus et al. (2022), and Tandoc et al. (2021)). It might be that in these contexts, analytical thinking cannot override participants' strong intuitions of right and wrong. Supporting this line of reasoning, our additional analyses in section D of the supplemental materials found an effect of analytical thinking on misinformation sharing but only for nonmoral stimuli.

While previous work on analytical thinking can, in certain contexts, explain why individuals eventually decide to share or not to share misinformation - and thus help develop countermeasures (e.g., accuracy nudges) - it still leaves open what makes individuals want to share misinformation in the first place. Our research could fill in this gap: People are motivated to share value-aligned, identity-affirming content. Our studies found that agreement and perceived moral alignment with a post increased sharing of misinformation, indicating motivational drivers, potentially further amplified by

moral-emotional responses to the moral framing of posts that are aligned with one's moral values. Some evidence for this idea comes from past research that found a link between emotional responses and analytical thinking on believing and sharing of misinformation (Li et al., 2022; Wang et al., 2020). Emotional responses are linked to increased sharing of misinformation while analytical thinking reduces misinformation sharing. Deliberation might be used, only for strong enough accuracy concerns or conversely weak value and identity-based motives, to "rethink" this impulse and thus not share a message if it is inaccurate. Notable work that integrates both cognitive and motivational drivers of misinformation is the integrative approach by Van Bavel et al. (2021). This model acknowledges the influence of multiple motivational drivers (e.g., accuracy or identity-based) on believing and sharing misinformation. Our findings can contribute to this work by informing on the limitations and constraints of different drivers of misinformation and their potential interplay.

Our findings also complement current literature on affective and motivational drivers of responses to (mis)information, which found that emotional responses, such as psychological discomfort (Susmann & Wegener, 2022), fear (Featherstone & Zhang, 2020), or anger (Thorson, 2016) influence processing, believing, and sharing of misinformation (Van Damme & Smets, 2014). Our work also confirms past findings that moral-emotional content elicits more engagement on social media platforms compared to non-moral-emotional content (Brady et al., 2017), and importantly, showcases that matching moral values and moral message content increases user engagement. Future work should investigate how far the effect of moral values and framing extends. For instance, past work has found that negative emotions, such as fear, anger, or anxiety, have a lasting effect on the perception of misinformation even after (successful) corrections (Cobb et al., 2013; Thorson, 2016) and might moderate partisanship effects. It would, therefore, be fruitful to investigate whether moral emotions (e.g., emotional responses from perceived moral transgressions) similarly impact perception of misinformation. This is especially

relevant considering that misinformation frequently features moral-emotional appeals (Ghanem et al., 2021; Lewandowsky et al., 2012; Yeo & McKasy, 2021).

Our work is also in line with past research that utilized values-based messages which appeal to core morality to influence individuals' attitudes and behaviors on a range of topics, such as vaccinations (Amin et al., 2017), mask-wearing (Kaplan et al., 2021), or climate change (Feinberg & Willer, 2013, 2019). Specifically, this line of work demonstrates that moral framing in line with recipients' moral values can be used to make specific misinformation more believable and increase sharing intentions. Thus, this work further extends the current literature on the effects of (moral) framing and re-framing on persuasion into the field of misinformation.

Our work comes with some limitations. Although our stimuli are arguably naturalistic – that is, we analyzed real-world Twitter data in Study 1 and used realistic posts (including real news headlines) in Study 2 and Study 3 – their presentation does not fully represent participants experience on social media platforms. Specifically, due to logistical study limitations (e.g., survey length), we showed participants the stimuli in Study 2 and Study 3 with no other (filler) content in-between, such as friends' messages or ads. Similarly, the stimuli shown may not reflect the type of content to which the participants are usually exposed (e.g., due to user-specific social media algorithms). This is relevant as "echo chambers" are frequently encountered on social media and most Americans see mostly ideologically concordant content online (Bakshy et al., 2015). Furthermore, this work focused on self-reported sharing intentions of social media posts on a specific social media platform (i.e., Twitter). Future work should expand the scope of the current study to investigate whether the effect of moral values and framing on belief and sharing of misinformation also translates into real-world behaviors, such as patterns of sharing information online or offline and especially changes in behaviors relevant to the content they've seen (e.g., voting patterns or health behavior).

Lastly, this work did not account for habits in social media sharing behavior. Social

media platforms are heavily invested in building a habitual user base as their automatic behavior is monetized and critical to their financial models (Anderson & Wood, 2021; Bayer et al., 2022; Docherty, 2020). Furthermore, social rewards (e.g., up-votes, likes, badges, notifications) which are powerful cues in habitual learning are integral parts of these platforms' designs (Bayer et al., 2022; Bayer & LaRose, 2018). Users then build habits of sharing content, including against one's beliefs, that elicits social rewards (especially attention) but is not necessarily accurate. This results in a significant proportion of misinformation online being shared by highly habitual users (Ceylan et al., 2023). In this context, future work should also investigate the role that moral values and message content plays in building sharing habits. Moral values and message content may shape sharing habits because content that aligns with recipients' values elicits more engagement (see this work or Brady et al. (2017) and Candia et al. (2022)). As such, habitual sharing might lead to sharing moral-emotional content that elicits engagement instead of accurate content, thus facilitating the sharing of misinformation. Pennycook and Rand (2021) further found that users' sharing intentions of false headlines were significantly higher than their accuracy ratings, potentially indicating habitual sharing of headlines independent of veracity and accuracy judgments (see Herrero-Diz et al. (2020) for a discussion of habitual sharing of news independent of veracity). In this way, cognitive factors, socio-affective factors and habits might tie into an integrated system of sharing and believing misinformation online.

Conclusion

Building on past work on socio-affective drivers of misinformation, moral psychology and re-framing, we demonstrated how targeting audiences' core moral values can increase their engagement with (mis)information online, thus facilitating its spread. Importantly, we find that it is not moral content per se that drives misinformation sharing but it is the *matching* of a message's moral content and an individual's moral values. Framing content

in line with target audiences' core values (e.g., Individualizing or Binding values) will increase sharing of misinformation, even when the underlying arguments, partisanship, and worldview are kept constant. This indicates that partisan divides in misinformation sharing might be explained through their underlying moral values and beliefs. Importantly, our findings are independent of cognitive drivers, such as analytical thinking and familiarity with the content, further highlighting the role of motivational drivers behind (mis)information. As such, this work advances our understanding of the psychological mechanisms by which moral values and message framing interact, thereby leading to more sophisticated models that integrate characterizations of messages' moral content and receivers' core moral values to predict the success of social cyber-attacks. Ultimately, this research may offer a novel important perspective on our post-truth world: simple, targeted re-framing of the same message contents can lead to higher acceptance and spread of misinformation.

Acknowledgments

References

Adger, W. N., Butler, C., & Walker-Springett, K. (2017). Moral reasoning in adaptation to climate change. *Environmental Politics*, *26*(3), 371–390.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, *6*(2), 1–33.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211–36.

Al-Saggaf, Y., & Nielsen, S. (2014). Self-disclosure on facebook among female users and its relationship to feelings of loneliness. *Computers in Human Behavior*, *36*, 460–468.

Amin, A. B., Bednarczyk, R. A., Ray, C. E., Melchiori, K. J., Graham, J., Huntsinger, J. R., & Omer, S. B. (2017). Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, *1*(12), 873–880.

Anderson, I. A., & Wood, W. (2021). Habits and the electronic herd: The psychology behind social medias successes and failures. *Consumer Psychology Review*, *4*(1), 83–99.

Aral, S., & Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, *365*(6456), 858–861.

Atari, M., Graham, J., & Dehghani, M. (2020). Foundations of morality in iran. *Evolution and Human Behavior*, *41*(5), 367–384.

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2022). *Morality beyond the weird: How the nomological network of morality varies across cultures* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/q6c9r

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

Bayer, J. B., Anderson, I. A., & Tokunaga, R. (2022). Building and breaking social media habits. *Current Opinion in Psychology*, 101303.

Bayer, J. B., & LaRose, R. (2018). Technology habits: Progress, problems, and prospects. *The psychology of habit*, 111–130.

Bossetta, M. (2018). The weaponization of social media: Spear phishing and cyberattacks on democracy. *Journal of international affairs*, *71*(1.5), 97–106.

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, *149*(4), 746.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Brinol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European review of social psychology*, *20*(1), 49–96.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of applied research in memory and cognition*, *8*(1), 108–117.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Candia, C., Atari, M., Kteily, N., & Uzzi, B. (2022). Overuse of moral language dampens content engagement on social media.

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), e2216614120. https://doi.org/10.1073/pnas.2216614120

Cheng, C., Barceló, J., Hartnett, A. S., Kubinec, R., & Messerschmidt, L. (2020). Covid-19 government response event dataset (coronanet v. 1.0). *Nature human behaviour*, *4*(7), 756–768.

Clarkson, E., & Jasper, J. D. (2022). Individual differences in moral judgment predict attitudes towards mandatory vaccinations. *Personality and Individual Differences*, *186*, 111391.

Cobb, M. D., Nyhan, B., & Reifler, J. (2013). Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychology*, *34*(3), 307–326.

Colliander, J. (2019). this is fake news: Investigating the role of conformity to other users views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, *97*, 202–215.

Darwish, K., Stefanov, P., Aupetit, M., & Nakov, P. (2020). Unsupervised user stance detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 141–152.

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin*, *40*(12), 1559–1573.

De Keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2020). Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin*, *46*(2), 204–215.

Dehghani, M., Atran, S., Iliev, R., Sachdeva, S., Medin, D., & Ginges, J. (2010). Sacred values and conflict over iran's nuclear program. *Judgment and Decision making*, *5*(7), 540.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., Vaisey, S., Iliev, R., & Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*, *145*(3), 366.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Docherty, N. (2020). Facebooks ideal user: Healthy habits, social capital, and the politics of well-being online. *Social Media+ Society*, *6*(2), 2056305120915606.

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

Erceg, N., Gali, Z., & Bubi, A. (2018). The psychology of economic attitudes–moral foundations predict economic attitudes beyond socio-demographic variables. *Croatian Economic Survey*, *20*(1), 37–70.

Fazio, L. K. (2020). Repetition increases perceived truth even for known falsehoods. *Collabra: Psychology*, *6*(1).

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993.

Featherstone, J. D., & Zhang, J. (2020). Feeling angry: The effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, *25*(9), 692–702.

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological science*, *24*(1), 56–62.

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, *41*(12), 1665–1681.

Feinberg, M., & Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, *13*(12). https://doi.org/10.1111/spc3.12501

Forgas, J. P., & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*, *44*(5), 1362–1367.

Ghanem, B., Ponzetto, S. P., Rosso, P., & Pardo, F. M. R. (2021). Fakeflow: Fake news detection by modeling the flow of affective information. *CoRR*, *abs/2101.09810*. https://arxiv.org/abs/2101.09810

Graham, J. (2013). Mapping the moral maps: From alternate taxonomies to competing predictions. *Personality and Social Psychology Review*, *17*(3), 237–241.

Graham, J., & Haidt, J. (2010). Beyond beliefs: Religions bind individuals into moral communities. *Personality and social psychology review*, *14*(1), 140–150.

Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach.

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

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, *363*(6425), 374–378.

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, *9*(3), 4.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), 110–119.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66.

Herrero-Diz, P., Conde-Jiménez, J., & Reyes de Cózar, S. (2020). Teens motivations to spread fake news on whatsapp. *Social Media+ Society*, *6*(3), 2056305120942879.

Hoover, J., Johnson, K., Boghrati, R., Graham, J., & Dehghani, M. (2018). Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, *4*(1).

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hurst, K., & Stern, M. J. (2020). Messaging for environmental action: The role of moral framing and message source. *Journal of Environmental Psychology*, *68*, 101394.

Jensen, M. (2018). Russian trolls and fake news: Information or identity logics? *Journal of International Affairs*, *71*(1.5), 115–124.

Jiang, X., Su, M.-H., Hwang, J., Lian, R., Brauer, M., Kim, S., & Shah, D. (2021). Polarization over vaccination: Ideological differences in twitter expression about covid-19 vaccine favorability and specific hesitancy concerns. *Social Media+ Society*, *7*(3), 20563051211048413.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural public policy*, *1*(1), 54–86.

Kaplan, J., Vaccaro, A., Henning, M., & Christov-Moore, L. (2021). Moral reframing of messages about mask-wearing during the covid-19 pandemic.

Kerr, J., Panagopoulos, C., & van der Linden, S. (2021). Political polarization on covid-19 pandemic response in the united states. *Personality and individual differences*, *179*, 110892.

Kivikangas, J. M., Fernández-Castilla, B., Järvelä, S., Ravaja, N., & Lönnqvist, J.-E. (2021). Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological Bulletin*, *147*(1), 55.

Koch, A. S., & Forgas, J. P. (2012). Feeling good and feeling truth: The interactive effects of mood and processing fluency on truth judgments. *Journal of Experimental Social Psychology*, *48*(2), 481–485.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). Combating fake news: An agenda for research and action.

Lee, S., Forrest, J. P., Strait, J., Seo, H., Lee, D., & Xiong, A. (2020). Beyond cognitive ability: Susceptibility to fake news is also explained by associative inference. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, *13*(3), 106–131.

Li, M.-H., Chen, Z., & Rao, L.-L. (2022). Emotion, analytic thinking and susceptibility to misinformation during the covid-19 outbreak. *Computers in Human Behavior*, *133*, 107295.

Low, M., Wui, M., & Lopez, G. (2016). Moral foundations and attitudes towards the poor. *Current Psychology*, *35*(4), 650–656.

MacGuill, D. (2021). Yes, Portland really did name a new bridge after Ned Flanders. Retrieved June 22, 2022, from https://www.snopes.com/fact-check/portland-ned-flanders-bridge/

Mackie, D. M., Worth, L. T., & Asuncion, A. G. (1990). Processing of persuasive in-group messages. *Journal of personality and social psychology*, *58*(5), 812.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C. D., et al. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835–3840.

Marietta, M. (2008). From my cold, dead hands: Democratic consequences of sacred rhetoric. *The Journal of Politics*, *70*(3), 767–779.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications*, *5*(1), 1–20.

Mueller III, R. S. (2019). Report on the investigation into russian interference in the 2016 presidential election. volumes i & ii.(redacted version of 4/18/2019).

Muric, G., Wu, Y., Ferrara, E., et al. (2021). Covid-19 vaccine hesitancy on social media: Building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance*, *7*(11), e30642.

Nyilasy, G. (2019). Fake news: When the dark side of persuasion takes over. *International Journal of Advertising*, *38*(2), 336–342.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, *115*(3), 999–1015.

Pennycook, G., & Rand, D. G. (2019a). Cognitive reflection and the 2016 us presidential election. *Personality and Social Psychology Bulletin*, *45*(2), 224–239.

Pennycook, G., & Rand, D. G. (2019b). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526.

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

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, *88*(2), 185–200.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388–402.

Pretus, C., Servin-Barthet, C., Harris, E., Brady, W., Vilarroya, O., & Van Bavel, J. (2022). The role of political devotion in sharing partisan misinformation.

Reimer, N. K., Atari, M., Karimi-Malekabadi, F., Trager, J., Kennedy, B., Graham, J., & Dehghani, M. (2022). Moral values predict county-level covid-19 vaccination rates in the united states. *American Psychologist*, *77*(6), 743.

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy*, *2*(1), 85–95.

Singh, M., Kaur, R., Matsuo, A., Iyengar, S., & Sasahara, K. (2021). Morality-based assertion and homophily on social media: A cultural comparison between english and japanese languages. *Frontiers in psychology*, 5081.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of personality and social psychology*, *88*(6), 895.

Skitka, L. J., & Morgan, G. S. (2014). The social and political implications of moral conviction. *Political psychology*, *35*, 95–110.

Skitka, L. J., & Mullen, E. (2002). The dark side of moral conviction. *Analyses of Social Issues and Public Policy*, *2*(1), 35–41.

Stan Development Team. (2021). RStan: The R interface to Stan. Retrieved September 9, 2021, from http://mc-stan.org/

Stroope, S., Kroeger, R. A., Williams, C. E., & Baker, J. O. (2021). Sociodemographic correlates of vaccine hesitancy in the united states and the mediating role of beliefs about governmental conspiracies. *Social Science Quarterly*, *102*(6), 2472–2481.

Sung, Y., Lee, J.-A., Kim, E., & Choi, S. M. (2016). Why we post selfies: Understanding motivations for posting pictures of oneself. *Personality and Individual Differences*, *97*, 260–265.

Sunstein, C. R. (2003). Moral heuristics and moral framing. *Minn. L. Rev.*, *88*, 1556.

Susmann, M. W., & Wegener, D. T. (2022). The role of discomfort in the continued influence effect of misinformation. *Memory & Cognition*, *50*(2), 435–448.

Swire, B., Ecker, U. K., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of experimental psychology: learning, memory, and cognition*, *43*(12), 1948.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, *50*(3), 755–769.

Tandoc, E. C., Lee, J., Chew, M., Tan, F. X., & Goh, Z. H. (2021). Falling for fake news: The role of political bias and cognitive ability. *Asian Journal of Communication*, *31*(4), 237–253.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, *11*(1), 99–113.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460–480.

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Valenzuela, S., Piña, M., & Ramrez, J. (2017). Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of communication*, *67*(5), 803–826.

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, *22*(3), 213–224.

Van Damme, I., & Smets, K. (2014). The power of emotion versus the power of suggestion: Memory for emotional events in the misinformation paradigm. *Emotion*, *14*(2), 310.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Voelkel, J. G., Malik, M., Redekopp, C., & Willer, R. (2022). Changing americans attitudes about immigration: Using moral framing to bolster factual arguments. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 73–85.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146–1151.

Wang, R., He, Y., Xu, J., & Zhang, H. (2020). Fake news or bad news? toward an emotion-driven cognitive dissonance model of misinformation diffusion. *Asian Journal of Communication*, *30*(5), 317–342.

Winkielman, P., Huber, D. E., Kavanagh, L., & Schwarz, N. (2012). Fluency of consistency: When thoughts fit nicely and flow smoothly. *Cognitive consistency: A fundamental principle in social cognition*, 89–111.

Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, *65*, 7–19.

Wu, T.-Y., & Atkin, D. (2017). Online news discussions: Exploring the role of user personality and motivations for posting comments on news. *Journalism & Mass Communication Quarterly*, *94*(1), 61–80.

Yeo, S. K., & McKasy, M. (2021). Emotion and humor as misinformation antidotes. *Proceedings of the National Academy of Sciences*, *118*(15), e2002484118.

Yin, L., Roscher, F., Bonneau, R., Nagler, J., & Tucker, J. A. (2018). Your friendly neighborhood troll: The internet research agencys use of local and fake news in the 2016 us presidential campaign. *SMaPP Data Report, Social Media and Political Participation Lab, New York University*.

Ziegler, C. E. (2018). International dimensions of electoral processes: Russia, the usa, and the 2016 elections. *International Politics*, *55*(5), 557–574.

Appendix A

Stimuli

Study 1

Table A1
Exemplary tweets showcasing moral framing and/or stance on COVID vaccinations and mandates

Topic	Description	Example
Nonmoral	Does not contain moral language	They really think mandating vaccines on airplanes is gonna sway the unvaccinated, lol. I guess Im gonna just drive...
Individualizing	Focused on individual rights and well-being	Under 12s are unvaccinated! We need to ensure all primary schools have safe air to prevent mass infections.
Binding	Focused on group preservation	Common law, natural law, God's law. I will never consent and am both disgusted and horrified by people's acquiescence... My body belongs to GOD!
Pro-vax	Endorsing COVID vaccines and mandates	Lets get vaxxed!
Anti-vax	Opposing COVID vaccines and mandates	No. Do Not get Vaccinated!

Study 2 & Study 3

Table A2
List of headlines selected for Study 2b and Study 3

#	Headline	True/False	Sentiment
1	Man Shoots Off His Left Ear Taking Selfies With Gun	False	Negative
2	Refugees Have 100 Times Greater Rate Of Tuberculosis Than National Average	False	Negative
3	Starbucks Is Giving Out Free Lifetime Passes On Its 44th Anniversary	False	Negative
4	Man Infects 586 People With HIV On Purpose, Plans On Infecting 2,000 More Before 2024	False	Negative
5	Trillionaires Now Exist!	False	Negative
6	America Is Now Reducing CO2 Emissions Much Faster Than Other Developed Countries	False	Positive
7	Black And White Wealth Gap Is Closing Fast	False	Positive
8	John Travolta Takes A New Wife After The Death Of Kelly Preston	False	Positive
9	Polar Bears Are Strong and Healthy Across Alaska	False	Positive
10	Taco Bell Reportedly Going Out of Business	False	Positive
11	Twitter Is Making A Dislike Button	True	Negative
12	Crocs is Giving Away Free Footwear to Healthcare Workers	True	Negative
13	Portland Named a New Bridge After The Simpsons Ned Flanders	True	Negative
14	State Department Charging Americans $2k For Flights Out Of Afghanistan	True	Negative
15	Whitest-Ever Paint Could Help Cool Heating Earth	True	Negative
16	Hole In The Ozone Layer Will Totally Heal Within 50 Years	True	Positive
17	An Invisible Sculpture Sold for $18K	True	Positive
18	A California Man Sued A Psychic For Allegedly Failing To Remove A Curse	True	Positive
19	Bluetooth Technology Was Named After A Viking King	True	Positive
20	Scientists Detect Cocaine In Freshwater Shrimp	True	Positive

Note. Headlines within each combination of veracity and sentiment are ordered by the criterion described in the Results section of Study 2a.

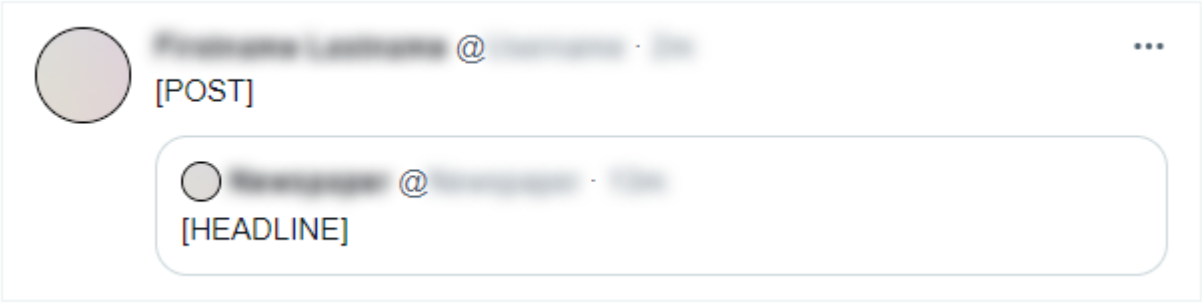
Appendix B

Questionnaire Items (Study 2a, 2b, 3)

Introduction

On the next few pages, you will see examples of social media posts that look somewhat like this:

Figure B1
Example stimulus presentation for the introduction



Each example contains a news headline a user has shared (here: [HEADLINE]) and a post the user has written about the news headline (here: [POST]). For this study, we are leaving out some information about the post (for example, who posted it and when). Please answer each of the questions as if you had come across the post while using social media (e.g., Twitter or Facebook). In total, you will answer questions about 5 posts.

Headline-Level

On this page, please focus on the headline the user has shared:

Figure B2
Example stimulus presentation for headline-level items

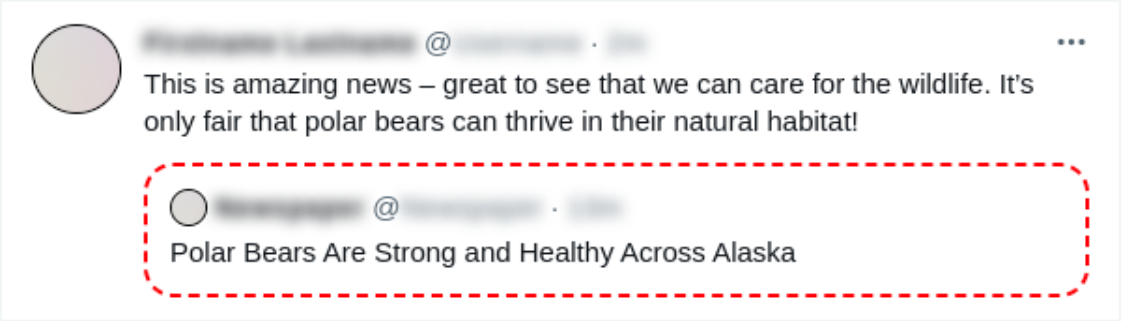


Table B1
In your opinion, the post the user has written about the headline is

Unbelievable	1	2	3	4	5	6	7	Believable
Uncontroversial	1	2	3	4	5	6	7	Controversial
Unsurprising	1	2	3	4	5	6	7	Surprising
Uninteresting	1	2	3	4	5	6	7	Interesting
Negative	1	2	3	4	5	6	7	Positive

Post-Level

On this page, please focus on the post the user has written about the headline:

Figure B3
Example stimulus presentation for post-level items

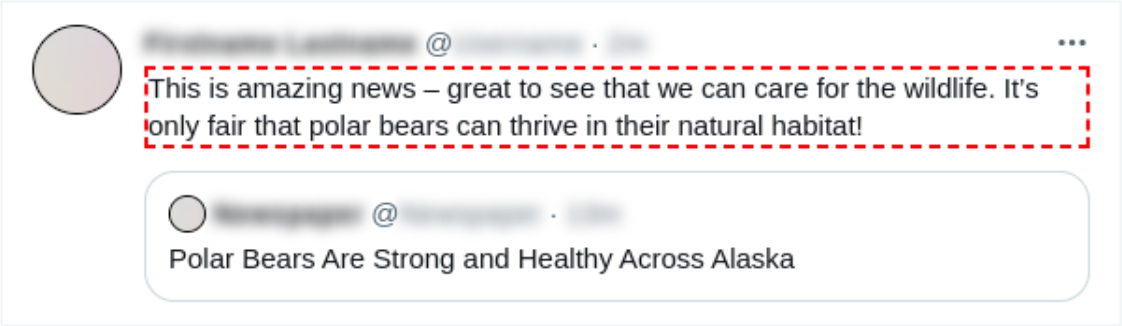


Table B2
In your opinion, the post the user has written about the headline is

Uncontroversial	1	2	3	4	5	6	7	Controversial
Unsurprising	1	2	3	4	5	6	7	Surprising
Uninteresting	1	2	3	4	5	6	7	Interesting
Negative	1	2	3	4	5	6	7	Positive

Table B3
How much do you agree or disagree with the post the user has written about the headline?

Strongly disagree	1	2	3	4	5	Strongly agree

Table B4
How much does the post the user has written about the headline align with your values?

Strongly opposed to my values	1	2	3	4	5	6	7	Strongly aligned with my values

On this page, please focus on the entire social media post:

Figure B4

Example stimulus presentation for whole-stimulus items

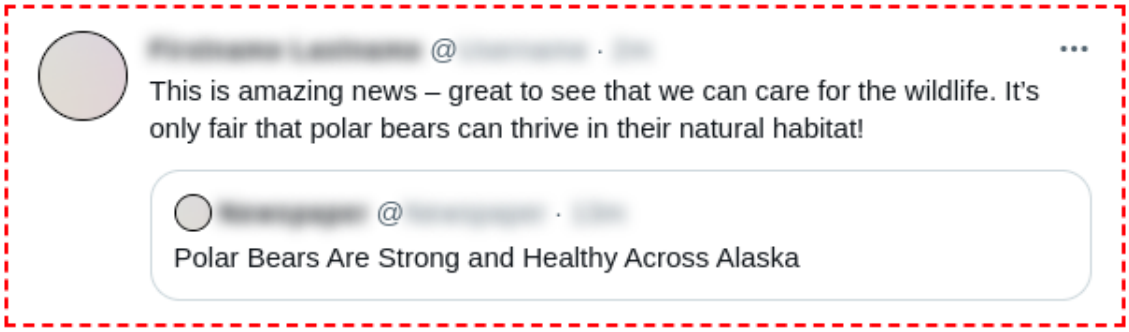


Table B5

If you came across this post, how likely would you be to ...

... share the post publicly on your social media feed (e.g., retweet or share on your Facebook)	1	2	3	4	5
... 'like' the post (e.g., Twitter or Facebook)?	1	2	3	4	5
... share the post privately in a private message, text message, or email?	1	2	3	4	5
... talk about the post or headline in an offline conversation?	1	2	3	4	5

Table B6

Why did you decide to share (or not to share) the previous post? (Only Study 3)

It just felt right/wrong without much thought	1	2	3	4	5	I carefully considered the information and implications
It matched my gut-feeling and intuitions	1	2	3	4	5	It matched my knowledge and experiences
It made me feel strongly	1	2	3	4	5	It made me very thoughtful
My social circle posts similar/different posts	1	2	3	4	5	I reflected on my own thoughts and opinions

Appendix C

Foundation-Level Analysis of Social Media Posts

For a more detailed analysis of the interaction between user stance and moral framing on user engagement (retweet count) in Study 1, we further investigated the individual moral foundations (Care, Fairness, Loyalty, Authority, and Purity)[8].

Regarding Individualizing foundations (Care & Fairness), we find, in line with our hypothesis, that Care and Fairness framing predicted more (1.4 times; 3.9 times) engagement for "pro-vax" users compared to "anti-vax" users ($\beta = 0.15, [0.02, 0.28]; \beta = 0.59, [0.41, 0.78]$). Regarding Binding foundations (Loyalty, Authority, and Purity), we found that, in line with our hypothesis, Authority and Purity framing predicted significantly more (3.31 times; 11.7 times) engagement for "anti-vax" users compared to "pro-vax" users ($\beta = -0.52, [-0.90, -0.10]; \beta = -1.07, [-1.68, -0.42]$). However, opposite to our hypothesis Loyalty predicted more (3.7 times) engagement for "pro-vax" compared to "anti-vax" users ($\beta = 0.57, [0.07, 1.08]$). Our findings regarding retweet count are therefore in line with our expectations regarding the underlying relationship of moral values and stances (ideology-aligned moral framing increases engagement) with the single exception of Loyalty framing. Past work has observed that Loyalty values and framing can be linked to increased vaccination rates (Reimer et al., 2022), which might explain this specific outlier. Importantly, however, this outlier was not large enough to influence the general finding that Binding framing facilitates "anti-vax" messages as shown in the main analyses.

We further investigated the effects of each individual foundation on favourite count. Regarding Individualizing foundations (Care & Fairness), we found in line with our hypothesis that Care and Fairness framing predicted more (4.2 times; 4.7 times) engagement for "pro-vax" compared to "anti-vax" users

[8] We trained another BERT-based classifier to detect the individual moral foundations analogue to the methods in Study 1. The classifier achieved a cross-validated F1 score of 0.71

($\beta = 0.62, [0.27, 0.98]; \beta = 0.67, [0.17, 1.17]$). Regarding Binding foundations (Loyalty, Authority, and Purity), we found, opposite to our hypothesis, that Authority framing predicted significantly more (46.8 times) engagement for "pro-vax" users compared to "anti-vax" users ($\beta = 1.67, [0.74, 2.62]$). Furthermore, we found no difference between "pro-vax" and "anti-vax" for Loyalty ($\beta = -0.97, [-2.29, 0.46]$) and Purity framing ($\beta = -0.71, [-2.33, 1.03]$). Thus, our findings are in line with our expectations for Individualizing framing but not for Binding framing, driven by a reversed effect of Authority.

Overall, the extended analysis shed light on which specific values drive the outlier observed in our main analyses in Study 1: Authority framing increasing liking of "pro-vax" tweets. These findings relate to some of the limitations discussed in the main paper. First, there could be conservatives (who endorse Binding values) that are not "anti-vax" and therefore endorse "pro-vax" arguments that include Authority framing. Similarly, there could be "pro-vax" arguments that confound with Authority language (e.g., "Respect experts/the law to keep everyone safe"; relates to both authority and care/harm avoidance). Depending on the distribution of conservatives and liberals that are shown the respective tweet, this might then lead to the observed effect, even if liberals engage less with tweets that have a mismatched moral framing. However, it should be noted that we observe this outlier only for likes but not for retweets, where the effect was in the predicted direction. This could be caused by user behavior differing for liking and sharing tweets, e.g., users could be less hesitant to like content that they would not share because it is less public.

Appendix D

Additional Analyses of Analytical Thinking

Study 3 replicated the results of Study 2b and tested whether lack of deliberation could be an alternative explanation. We tested whether an alignment of moral values and framing distracted participants from post accuracy and plausibility (via reducing deliberation), leading to more sharing of misinformation. We did not find that deliberation of posts mediated the effect of aligning a participant's moral values and a post's framing. To further strengthen these findings, we directly tested whether deliberation reduced sharing of false (vs true) news by fitting four additional linear regression models.

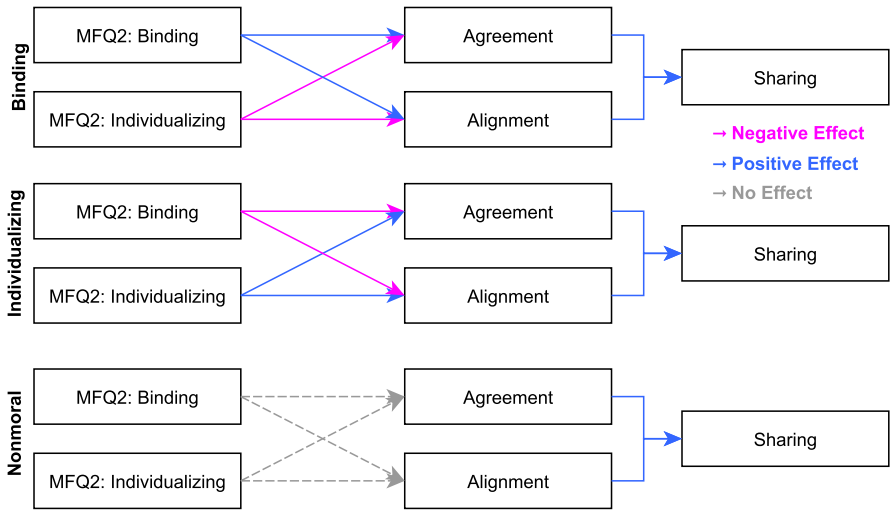
First, we fitted model M7 that predicted sharing intentions as a function of analytical thinking, headline veracity and their interaction. This model tested whether analytical thinking (CRT-2) reduced sharing of false (vs true) news. We found no effect for this relationship ($\beta = 0.03, [-0.04, 0.08]$). Second, we fit model M8 that predicted sharing intentions as a function of deliberating over sharing a post, headline veracity and their interaction. This model tested whether deliberation, over each sharing behavior, reduced sharing of fake (vs true) news. We, again, found no evidence for this relationship ($\beta = 0.01, [-0.09, 0.11]$). Furthermore, these models did not predict sharing intentions more accurately than the null model ($\Delta_{ELPD} = -2.61, SE = 8.52, z = -0.31; \Delta_{ELPD} = 7.70, SE = 9.04, z = 0.85$), which predicted sharing intentions as a function of random intercepts for participants, headlines, and posts. Third, we fitted model M9 that predicted sharing intentions as a function of headline veracity, analytical thinking (CRT-2), headline believability and their interaction. This model tested whether analytical thinking increased accuracy and plausibility considerations, as argued by past work (Pennycook & Rand, 2019c). If analytical thinking results in participants estimating and considering plausibility, then there should be a positive interaction of CRT and headline believability, meaning that analytical thinkers should have higher plausibility concerns than "lazy" thinkers (Pennycook & Rand, 2019c).

We found no effect for this relationship ($\beta = -0.03, [-0.08, 0.02]$). Fourth, focusing only on nonmoral stimuli, we fitted model M10 that predicted sharing intentions as a function of analytical thinking, headline veracity, and their interaction. This model investigated whether the failure to replicate past findings of analytical thinking reducing misinformation sharing was caused by most of our stimuli being moralized (two thirds). It could be that for moral-emotional stimuli accuracy concerns were superseded by participants' intuitions of right and wrong. Supporting this idea, we find that, for nonmoral stimuli, analytical thinking reduces sharing of misinformation$\beta = -0.10, [-0.17, -0.02]$) but not true information ($\beta = -0.04, [-0.14, 0.05]$).

We further replicated the mediation analysis from Study 2, which mediated the effect of matching a participant's moral values and a post's moral framing on sharing intentions via agreement and moral alignment with the post. Similar to Study 2b, we compared across the three moral framing conditions the total indirect effects of participants' endorsement of Binding and Individualizing values on sharing intentions via their ratings of how much they agreed with the post, how much the post aligned with their moral values, and the interaction of the two ratings, while controlling for headline veracity and, additionally, headline familiarity. Figure D1 provides an overview of the observed relationships. Supporting the original Hypothesis 2 in Study 2b, we found that participants' endorsement of Binding values had a positive indirect effect on sharing intentions in the Binding framing condition ($\beta = .21, [.15, .27]$) but a negative indirect effect in the Individualizing framing condition ($\beta = -.10, [-.15, -.05]$). Furthermore, participants' endorsement of Individualizing values had a positive indirect effect on sharing intentions in the Individualizing framing condition ($\beta = .21, [.15, .28]$) but a negative indirect effect in the Binding framing condition ($\beta = -.08, [-.14, -.03]$). Lastly, participants' endorsement of Binding ($\beta = .00, [-.05, .06]$) and Individualizing ($\beta = .03, [-.02, .09]$) values had no indirect effect in the nonmoral framing condition. These findings, again, support the idea of motivational drivers of message sharing.

Figure D1

Replication of the mediation in Study 2b: Effect of matching moral values and framing via perceived agreement and alignment with the post



Appendix E

Additional Analyses of Order Effects

In our studies 2 & 3, we presented participants repeatedly with stimuli and items to rate said stimuli (e.g., believability, agreement, etc). Participants were then asked for their sharing intentions. Thus, participants might have learned after the first trial that they will have to indicate their sharing intentions for the presented social media posts at each trial. In the following iterations participants might have decided about their sharing intentions during post presentation (before all ratings) and this could have influenced their subsequent stimuli ratings (i.e., rating to justify their sharing intentions; e.g., rate a headline as more believable if they want to share the post). Therefore, as a robustness check, we conducted an additional analysis to test for potential order effects, that is whether the effect of our ratings on sharing intentions increased over trial iterations. We ran the respective models (M1, M3, M6) again while controlling for stimulus order but found no significant interactions of stimulus order and main predictors (M1: $\beta_{familiar:order} = -0.00, [-0.02, 0.01]$), M3: $\beta_{agreement:order} = 0.00, [-0.03, 0.02]$; M6: $\beta_{order:framing1:individualizing} = 0.00, [-0.02, 0.02]$, $\beta_{order:framing2:binding} = 0.01, [-0.02, 0.03]$).

Appendix F

Additional Models

Study 1

Table F1

Comparison of models estimating engagement (favorite count) in Study 1 as a function of various predictor variables

			z		
Model	Description	R^2	M0	M1	M2
M0	Stance	0.05	-	-2.56	1.55
M1	Moral framing (all foundations)	0.04	2.56	-	4.75
M2	Moral framing (all foundations) & stance interaction	0.05	-1.55	-4.75	-

Note. R^2 *is a Bayesian analogue to the proportion of within-sample variance explained by a model (not considering varying effects).* z *is the difference in out-of-sample prediction accuracy between two models divided by its standard error* ($z = \Delta_{ELPD}/SE$).

Table F2

Comparison of models estimating engagement (retweet count) in Study 1 as a function of various predictor variables

			z		
Model	Description	R^2	M0	M1	M2
M0	Stance	0.10	-	1.84	-3.28
M1	Moral framing (all foundations)	0.10	-1.84	-	-4.93
M2	Moral framing (all foundations) & stance interaction	0.10	3.28	4.93	-

Note. R^2 *is a Bayesian analogue to the proportion of within-sample variance explained by a model (not considering varying effects).* z *is the difference in out-of-sample prediction accuracy between two models divided by its standard error* ($z = \Delta_{ELPD}/SE$).

Study 2b

Table F3

Effect sizes for Model 4 when controlling for headline veracity

Values - Condition	β[CI]	$\Delta\beta_{no\ control}$
Binding - Binding	.26 [.16, .36]	0.0
Binding - Individualizing	.14 [.04, .24]	0.0
Binding - Nonmoral	.20 [.10, .30]	0.0
Individualizing - Binding	.07 [-.01, .14]	0.0
Individualizing - Individualizing	.23 [.16, .26]	0.0
Individualizing - Nonmoral	.14 [.06, .21]	0.0
Proportionality - Binding	.00 [-.09, .09]	0.0
Proportionality - Individualizing	-.05 [-.14, .04]	0.0
Proportionality - Nonmoral	-.03 [-.12, .07]	0.0
$\Delta\beta_{Binding:Binding-Individualizing}$	.12 [.03, .21]	0.0
$\Delta\beta_{Binding:Binding-Nonmoral}$	.06 [-.04, .15]	0.0
$\Delta\beta_{Individualizing:Individualizing-Binding}$	.17 [.09, .24]	0.01
$\Delta\beta_{Individualizing:Individualizing-Nonmoral}$	.10 [.01, .18]	0.0

Note. Table shows the effect sizes for a given moral value in a given framing condition, as well as the difference in effect sizes for a moral value across framing conditions (last 4 rows). Table shows that the reported effect sizes in Study 2b hold when controlling for headline veracity.