Targeting Audiences' Moral Values Shapes Misinformation Sharing

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, research materials, and preregistrations are available under https://osf.io/z25tc/?view_only=0141845d12024a2cbdbd0f71f77f23a8. A preprint of an early version has been archived at https://osf.io/preprints/psyarxiv/ztq2k. The authors declare no conflicts of interest. This project was funded in part by National Science Foundation SaTC (Award No: 2140473) and Defense Advanced Research Projects Agency (DARPA) HR001121C0165. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

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 moral values facilitate the spread of misinformation? We investigate this question in three behavioral experiments, two of which were preregistered ($N_{1a} = 615$; $N_{1b} = 505$; $N_2 = 533$), in conjunction with analyzing real-world Twitter data (N = 20,235; 809,414 tweets) on misinformation related to COVID vaccinations. First, we investigate how 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 but find no effect of analytical thinking. Furthermore, we find that moral alignment facilitates sharing misinformation more so than true information. Next, we use natural language processing to determine messages' moral framing and political ideology on COVID vaccinations and mandates. We find that an alignment of moral framing and ideology facilitates the spread of COVID vaccination misinformation. 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

Public Significance Statement

Our research reveals that aligning online messages with audiences' core moral values through framing strategies is a powerful mechanism to influence (mis)information sharing. This effect can supersede cognitive factors, such as analytical thinking ability, and emphasizes the significant impact of motivational factors, like alignment with core values, in spreading misinformation. Our findings underscore the urgent need for effective countermeasures against potentially targeted misinformation campaigns to mitigate the societal risks posed by their unchecked spread.

Manuscript word count: 9,424

Targeting Audiences Moral Values Shapes Misinformation Sharing

The prevalence of misinformation poses an imminent threat to our society. Misinformation, here defined as information that turns out to be incorrect or unsubstantiated (Ecker et al., 2022), has been linked to increased political polarization, altering perception of public figures and political issues, as well as undermining trust in key democratic institutions (Allcott & Gentzkow, 2017; Ciampaglia et al., 2018; Hochschild & Einstein, 2015), particularly in combination with their rapid and large-scale spread on social media networks (D. Lazer et al., 2017; D. M. Lazer et al., 2018). Given the substantial impact that misinformation can have on society, it is essential to determine contributing factors to the increasing spread of misinformation.

From a general perspective on information sharing, individuals are motivated to share information, such as news articles or personal accounts of current and past events, for diverse reasons. These motivations include expressing themselves by reflecting on their emotions, values, and worldview (Oh & Syn, 2015), assisting others, such as supporting communities during natural disasters or for altruistic purposes (Dong et al., 2021; Oh & Syn, 2015; Osatuyi, 2013), seeking personal gains or reputation (Erickson, 2011; Oh & Syn, 2015), and the intrinsic inclination to share information to facilitate cooperation and resource-sharing (see the application of Social Exchange Theory in information sharing, e.g., Osatuyi, 2013). Online sharing, particularly on social media platforms, further enables real-time information dissemination with minimal effort requirements, and the ability to reach vast audiences (Oh & Syn, 2015; Osatuyi, 2013). Thus, motivations for sharing information often pertain to helping others and managing ones reputation. Yet, misinformation remains prevalent on social media, despite the fact that it often contradicts widely accepted facts or exhibits exaggerations (Ecker et al., 2022; Pennycook et al., 2021).

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, 2019b). 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 moral language could not only persuade users to believe misinformation but also 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 interaction 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).

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). 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). Importantly, the specific language and framing used influence the acceptance of information beyond political beliefs conveyed in the very same message, i.e., the message being pro-Democrat or pro-Republican (Feinberg & Willer, 2013, 2015; Wolsko et al., 2016). This suggests that messages can gain efficacy in part by resonating 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). 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 might be facilitated by moral language in order to mitigate these severe consequences.

Our work, thus, seeks to elucidate how moral framing interacts with audiences' values to facilitate the spread of online messages, particularly misinformation due to its often politicized and moralized nature (Brady et al., 2017; Crockett, 2017). This complements investigations into online manipulation through 'microtargeting,' particularly relevant because modern AI tools such as large language models (LLMs), can now automatise large-scale creation and spreading of personalized messages (Simchon et al., 2024). Specifically, this work investigates the effect of matching message framing and individuals' values on the spread of (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 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 which 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, 2019b; Van Bavel & Pereira, 2018; Winkielman et al., 2012).

We hypothesize that messages that align with audiences' core moral values will be more effective than those that are misaligned or do not target core moral values, and that this effect will hold for messages containing misinformation. We expect that, in the U.S., misinformation framed around Binding values is more effective in specifically persuading political conservatives, and conversely, misinformation that relies on Individualizing framing is 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 whereas Binding values correlate positively with political conservatism (Kivikangas et al., 2021).

---

[1] Note, that recent research suggests that these superordinate categories might be specific to Western cultures (Atari et al., 2020)

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, 2019b, 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 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 how moral values motivate behavior: Dehghani et al. (2016) and Kahan et al. (2017)) and these intuitions can supersede accuracy concerns. In that case, there should not be an effect of deliberation on (mis)information sharing.

To investigate the relationship between moral framing and responses to shared content, we conduct two sets of studies. First, we develop a paradigm that allows us to directly test how matching of moral framing with the audiences' moral values affects responses to shared social media content in a controlled experimental setting. We then use this paradigm in two pre-registered studies to confirm the proposed effects and to shed light on the underlying psychological mechanisms. Second, we analyze real-world social media data (Twitter) containing rumors and misinformation about COVID-19 vaccinations and mandates to investigate the relationship between a message's moral framing and the sender's political ideology on engagement and test whether the effect of moral alignment holds in naturalistic online data. Together, our work provides additional insight into how the alignment of moral values and message framing may contribute to the spread of (mis)information.

While this work is centered on general misinformation instead of specific types of misinformation, such as disinformation (i.e., misinformation that is spread intentionally), it is still relevant for the latter. 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; D. Lazer et al., 2017; Nyilasy, 2019; Vosoughi et al., 2018). Most notably, malicious actors have manipulated narratives, amplified inflammatory messages, and 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 that *could* be used by malicious actors to spread misinformation in order to mitigate vulnerabilities to such attacks.

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data, analysis code, and research materials are available at URL. Machine learning models were trained and applied using `Python`, version 3.9 (Van Rossum & Drake, 2009) via the packages `TensorFlow` (Martín Abadi et al., 2015) and `Keras` (Chollet et al., 2015). Statistical analyses were conducted in `R` (R Core Team, 2022) using the "brms" library (Bürkner, 2017, 2018). Studies 1 and 2 were pre-registered on OSF under https://osf.io/f7r8d?view_only=7e4b1b5e3c574be6848664235fbd41ca and https://osf.io/69p4e?view_only=56f3d83e0fd6434989a7c7ead4ca0f40 respectively.

Study 1

In Study 1, we test our hypotheses about the relationships between moral framing, moral values, and responses to shared social media content. Specifically, we conduct two experiments to (1) develop a paradigm for studying how moral framing affects responses to shared social media content (Study 1a) and (2) use this paradigm to test our hypotheses (Study 1b). Study 1 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. 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 specifically *because* they agreed with the post and *because* it aligned with their moral values (Hypothesis 2).

Study 1a

In Study 1a, we developed a set of stimuli consisting of social media posts containing either true or false news headlines. These posts were framed to either align 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 Prolific, 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 the Prolific prescreening. 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 Prolific prescreening questionnaire. Provided that participants' answers matched their pre-screening responses, 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 values 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 1a, participants in Study 1a also completed a subset of additional measures used in Study 1 to facilitate

exploratory analysis and piloting. See an overview of our measures in section 2 of the Supplementary Information (SI).

Results

To select stimuli for Study 1, 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. We calculated an index of Binding values by averaging a participant's endorsement of the Loyalty, Authority, and Purity foundations and an index of Individualizing values by averaging a participant's endorsement of the Care and Equality foundations. We selected those posts as stimuli for Study 1b for which participants perceived alignment between the posts and their own values was highly correlated with the extent to which they themselves held one value (e.g., Binding) but not the other (e.g., Individualizing). Thus, a post framed using Binding values evoked higher alignment among participants with Binding values but not with Individualizing values and vice versa a post framed using Individualizing values evoked higher agreement among participants with Individualizing values but not with Binding values.

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 S1 in the SI). This way, Study 1a resulted in a paradigm that facilitates the investigation of how moral framing affects responses to shared social media content.

Study 1b

In Study 1b, 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

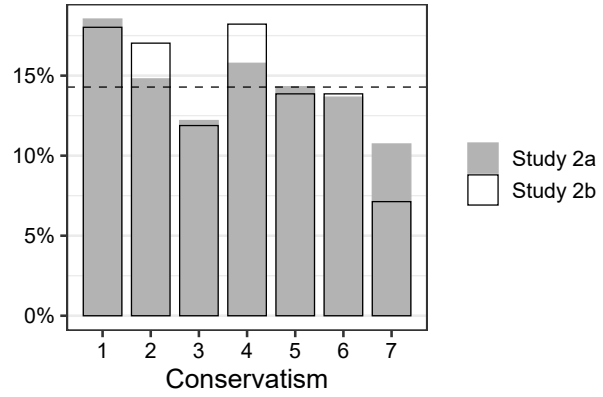
We preregistered the sample size as well as all hypotheses, inclusion/exclusion criteria, statistical models, measures, and manipulations (https://osf.io/f7r8d/?view_only=7e4b1b5e3c574be6848664235fbd41ca). We made all materials, data, and analysis scripts available online (https://osf.io/z25tc/?view_only=0141845d12024a2cbdbd0f71f77f23a8).

Participants. We recruited 641 U.S. American Twitter users from the Prolific subject pool who, according to Prolific, 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 the stimuli creation and selection (see section in the SI for details). 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. We further tested whether the limited number of conservatives would provide sufficient power. Note that power analyses in Bayesian statistics are not common, however, we conducted a simulation-based approach similar to Elsey (2021). We simulated the data in line with our experimental design, that is responses nested in headlines, posts, and users, with each user responding to 6 different headlines and tested for detecting a small effect size

($\beta = 0.1$) using weak, uninformed priors (see average effect sizes when targeting audiences' characteristics in: Joyal-Desmarais et al., 2022). The analyses revealed sufficient power for 500 participants ($power > 0.9$). See the code for the analyses in the OSF repository. As Figure 1 shows, our sample spanned the whole spectrum of political orientation.

Figure 1

Distribution of political orientation across samples

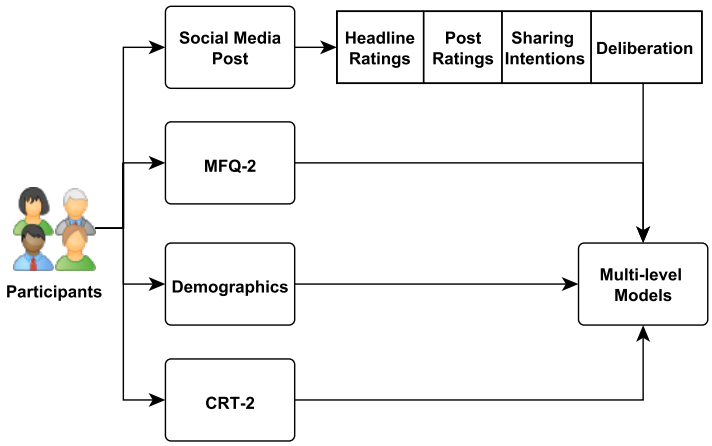


Note. The dashed line shows proportions expected under a uniform distribution.

Procedure. We used a planned missingness design that allowed both within-subject and between-subject comparisons. In total, we included 2 (headline: true, false) × 2 (post: positive, negative sentiment) × 5 = 20 news headlines selected during piloting of the paradigm (see Table S1 and section in the SI). 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 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 S1 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

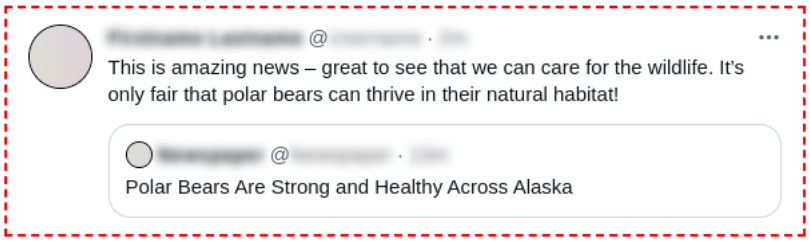
section 1 and 2 of the SI.

Figure 2
Illustration of Study flow



Note. Participants are presented with a social media post containing a news headline and text. Participants give headline-level and post-level ratings, and indicate their sharing intentions and deliberation over sharing (Study 2). Lastly, participants complete the MFQ-2, CRT-2 (Study 2), and demographic questions.

Figure 3
Exemplary stimulus presentation (shared news headline).



Note. Presented social media post contain a headline (bottom) and text about the headline (top).

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 about the headline (1 = *strongly disagree*, 5 = *strongly agree*) and how much the post 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 investigated the factors influencing sharing intentions while also comparing the predictive power of different types of predictors (e.g., do features of the headline predict sharing intentions more than features of the post or of the participant). We ran a series of Bayesian multilevel linear regression models that estimated participants' z-standardized sharing intentions as a function of various predictor variables:

Model 0, our baseline model, 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, investigating the effect of headline-level features, 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, investigating the effect of headline-level

features, 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, investigating the effect of perceived agreement with the post's contents, 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. Our main Model 4, investigating the effect of moral alignment, 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 values, the dummy-coded framing of each post, and the interaction between the two. Model 5, investigating the effect political ideology, mirrored Model 4 but included participants' z-standardized conservatism instead of their endorsement of moral values. Lastly, we ran a multilevel mediation model to estimate the indirect effects of moral values on sharing intentions via ratings of agreement and perceived moral alignment with each post. See Tables 1 for an overview of the model descriptions, and Tables S9 - S10 in the SI for the specific `R` formulas.

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

Table 1
Overview of Bayesian Multilevel Linear Regression Models

Model	Predictors	Coefficients
M0	Baseline without predictors	Random intercepts for posts, headlines, participants
M1	Headline-level ratings: believable, controversial, surprising, interesting, positive	Fixed effect of average ratings; Fixed and random effects of deviations from average
M2	Post-level ratings: controversial, surprising, interesting, positive	Fixed effect of average ratings; Fixed and random effects of deviations from average
M3	Agreement & alignment with post content	Fixed effect of average ratings and their interactions; Fixed and random deviation from average
M4	Participant's moral values, post's moral framing	Fixed and random effects of moral values, posts' moral framing and interaction
M5	Participant's political ideology, post's moral framing	Fixed and random effects of conservatism, moral framing and interaction
Mediation	Perceived agreement/alignment with a post mediates the effect of moral alignment	Fixed and random effects of moral values, post's moral framing, agreement with post and interactions

Note. Table provides an overview of the specification of the models in this Study. All models include a random intercept for posts, participants, and headlines.

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.

Results

Preregistered Analyses. Table 2 compares the models' out-of-sample prediction accuracies to each other. 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 ($\Delta_{ELPD} = 59.11, SE = 16.73, z = 3.53$). 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

Table 2
Comparison of preregistered models estimating sharing intentions as a function of various predictor variables

			z					
Model	Predictors	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 Values & Framing	.08	3.53	-8.83	-12.41	-8.30	-	2.89
M5	Political Orientation & Framing	.02	1.16	-11.73	-15.36	-11.00	-2.89	-

Note. R^2*, here, is its Bayesian counterpart.* z *is the standardized difference in out-of-sample prediction accuracy between each model (*$z = \Delta_{ELPD}/SE$*).*

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 ($\beta = 0.00, [-0.09, 0.09]; \beta = -0.05, [-0.14, 0.04]; \beta = -0.03, [-0.12, 0.06]$)[2]. See Table 3 an overview of the effect sizes of moral alignment vs misalignment and neutral posts on sharing intentions.

Model 4 predicted sharing intentions more accurately than Model 5 ($\Delta_{ELPD} = 48.81, SE = 16.89, z = 2.89$), which predicted sharing intentions as a function of political orientation instead of moral values. Taken together, these findings emphasize the facilitatory effect of targeting people's moral values on sharing (mis)information. However, models that estimated sharing intentions as a function of headline-level ratings (M1; $z = 8.83$), post-level ratings (M2; $z = 12.41$), or post-level alignment and agreement ratings (M3; $z = 8.30$) made more accurate out-of-sample predictions than Model 4. Across Model 1–3, 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

[2] 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 S18 in the SI for effect sizes when controlling model 4 for headline veracity.

Table 3
Effect of participant values on sharing across framing conditions (Main model M4)

	Aligned vs Misaligned	Aligned vs Neutral
Binding Values	0.12 [0.03, 0.21]	0.06, [-0.03, 0.15]
Individualizing Values	0.16 [0.09, 0.24]	0.10 [0.01, 0.18]

Note. Table shows the difference in the effect of moral alignment vs misalignment or non-moral (neutral) framing on sharing intentions. Table shows a significant effect of moral alignment vs misalignment for both values and a significant effect of moral alignment vs neutral posts for Individualizing values.

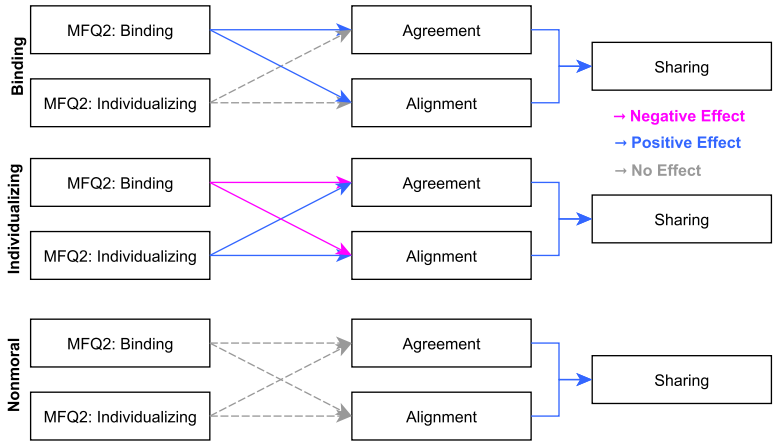
the post to be interesting (M2: $\beta = 0.34, [0.31, 0.38]$) and positive (M2: $\beta = 0.13, [0.09, .16]$); agreed with the post (M3: $\beta = 0.16, [0.11, 0.21]$), and considered the post to align with their moral values (M3: $\beta = 0.22, [0.17, 0.28]$). These findings were, perhaps, not surprising 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 interactions: Al-Saggaf and Nielsen (2014), Sung et al. (2016), and Wu and Atkin (2017)). Nevertheless, our findings show that (perceived) alignment of shared content and participant values has a significant facilitating effect on sharing intentions.

To test Hypothesis 2, that agreement and alignment with a post mediate the effect of morally aligned framing on sharing intentions, we estimated a Bayesian multilevel mediation model and compared 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 their interaction, while controlling for headline veracity. Figure 4 provides an overview of the observed relationships. Supporting Hypothesis 2, we found that 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 ($\beta = -.12, [-.20, -.04]$). Furthermore, 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 ($\beta = -.03, [-.10, .04]$). Lastly, 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 a positive mediation (blue color) for a match of moral framing and moral values, and no effect (grey) or a negative effect (red) for a mismatch.

Additional Analyses. Considering that headline-level ratings exhibited some of the strongest and most consistent effects on sharing intentions, and given that model M1 (headline ratings) outperformed model M4 (moral alignment), whether moral alignment has explanatory power above and beyond stimuli-features remains an open question. Thus, we fit an additional model $M1_{total}$ that combines headline-level ratings, moral values, and moral framing. We find that the effect of moral alignment persists even when controlling for the influence of headline-level ratings. Model $M1_{total}$ predicted sharing intentions more accurately than Model M1 ($\Delta_{ELPD} = 42.61, SE = 14.89, z = 2.86$), which did not include moral alignment, indicating that moral alignment adds explanatory power above and beyond headline-level ratings. Consistent with the main hypothesis, we find that participants' endorsement of Binding values predicted greater sharing intentions in the

Binding framing condition ($\beta = 0.23, [0.14, 0.32]$) than in the Individualizing framing condition ($\beta = 0.12, [0.04, 0.21]; \Delta\beta = 0.11, [0.02, 0.19]$) and, to a lesser extent, in the nonmoral framing condition ($\beta = 0.18, [0.09, 0.26]; \Delta\beta = 0.05, [-0.03, 0.14]$). 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 ($\beta = 0.17, [0.11, 0.24]$) than in the Binding framing condition ($\beta = 0.05, [-0.02, 0.12]; \Delta\beta = 0.13, [0.05, 0.20]$) and in the nonmoral framing condition ($\beta = 0.10, [0.03, 0.17]; \Delta\beta = 0.07, [+0.00, 0.15]$). 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).

Note that in all previous analyses, we tested the effects of different predictors on sharing intentions while controlling for the veracity of the headline. That is, we determined whether the observed effect of moral alignment is independent of stimuli veracity. However, to make inferences about whether moral alignment specifically facilitates misinformation beyond other online content, we refitted model M4 with a veracity and moral alignment interaction (model $M4_{veracity}$). We then analyzed whether the effects of moral alignment (and misalignment) differed between true and false content to more directly answer the question: "Does targeting audiences' core values facilitate the spread of misinformation?"

We find that the effect of moral alignment differed for true vs false content. That is, for misinformation, participants' endorsement of Binding values predicted greater sharing intentions in the Binding framing condition than in the Individualizing framing condition ($\Delta\beta = 0.12, [0.03, 0.21]$) and, to a lesser extent, in the nonmoral framing condition ($\Delta\beta = 0.06, [-0.04, 0.15]$), analogous to the previous findings of model M4. For true information however, participants' endorsement of Binding values did not predict greater sharing intentions in the Binding framing condition than in the Individualizing framing condition ($\Delta\beta = 0.06, [-0.07, 0.19]$) or in the nonmoral framing condition

Table 4
Effect of participant values on sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.12 [0.03, 0.21]	0.06, [-0.07, 0.19]	0.06, [-0.04, 0.15]	-0.03, [-0.08, 0.09]
Individualizing Values	0.16 [0.09, 0.24]	0.13 [0.09, 0.24]	0.09, [+0.00, 0.18]	0.06, [-0.06, 0.18]

Note. Table shows the difference in the effect of moral alignment vs misalignment or non-moral (neutral) framing on sharing intentions for both false and true posts. Table shows that the effect of moral alignment is generally stronger for false posts than for true posts and that for true posts the effect of moral alignment is only significant for Individualizing values and framing (aligned vs misaligned but not aligned vs neutral).

($\Delta\beta = -0.03, [-0.08, 0.09]$). In other words, participants endorsing Binding values showed higher sharing intentions for sharing misinformation (but not true information) framed with Binding values (aligned) than posts with Individualizing values (misaligned). Likewise, for misinformation, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition than in the Binding framing condition ($\Delta\beta = 0.16, [0.09, 0.24]$) and in the nonmoral framing condition ($\Delta\beta = 0.09, [+0.00, 0.18]$). For true information, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition than in the Binding framing condition ($\Delta\beta = 0.13, [0.09, 0.24]$) and, to a lesser extent, in the nonmoral framing condition ($\Delta\beta = 0.06, [-0.06, 0.18]$). In other words, participants with Individualizing values had greater sharing intentions for misinformation framed with Individualizing values (aligned) than nonmoral posts (neutral) and posts with Binding values (misaligned). For true information, this effect was dampened, and participants had greater sharing intentions only for posts framed with Individualizing values (aligned) vs Binding (misaligned). See Table 4 for an overview of the effect sizes of moral alignment vs misalignment and neutral posts on sharing intentions. Note that the

effect sizes of moral alignment were, across all conditions, lower for true information compared to misinformation even when the effects were significant (e.g., Binding vs Individualizing framing for participants with Individualizing values), indicating a generally higher sensitivity of misinformation to moral alignment.

Summary

The results of Study 1 demonstrate 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 whether the post contains true or false content. More detailed analyses further demonstrated that the influence of moral alignment (compared to non-moral and misalignment) predominantly drives misinformation sharing. This effect is not observed for true information sharing, where the facilitating impact of alignment either diminishes or disappears. Importantly, we also found that a match of framing and values predicts sharing intentions more accurately than other related variables, such as political ideology.

Study 2

The results of Study 1 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 information sharing by distracting participants from deliberating over post veracity or plausibility. 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 about the moralized posts, and this motivation could supersede accuracy concerns. In this case, there should not be an effect of deliberation on sharing intentions.

We first replicate Study 1, 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 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 self-reported 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[3] and made all materials, data, and analysis scripts available online[4].

Method

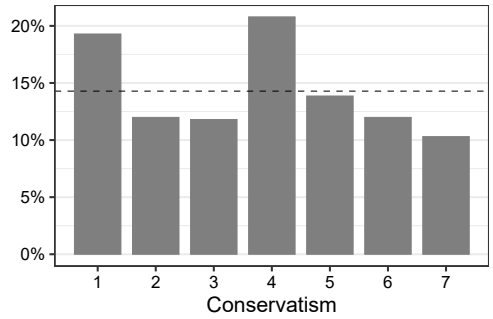
Participants. We recruited 676 U.S. American Twitter users from the Prolific subject pool who, according to Prolific, were U.S. residents, used Twitter at least once a month, posted on Twitter at least 1–3 times in the last 12 months, and who had not participated in Study 1a or 1b. We excluded participants who failed at least one of three attention checks or whose responses conflicted with the Prolific prescreening. 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

[3] Preregistration Link

[4] OSF Repository Link

moderate, and 178 identified as liberal. As Figure 5 shows, our sample spanned the spectrum of political orientation.

Figure 5
Distribution of political orientation across samples



Note. The dashed line shows proportions expected under a uniform distribution.

Procedure. We used the same planned missingness design from Study 1 that allowed both within-subject and between-subject comparisons. We included the same 2 (headline: true, false) × 2 (post: positive, negative sentiment) × 5 = 20 news headlines selected in Study "1a" (Table S1 in the SI) and used in Study 1. 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.

Measures. We collected the same post and headline-level ratings as in Study 1 (see Figure 2). To increase the 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 the 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.

Analysis Strategy. We replicated the five multilevel models from Study 1 that estimated participants' sharing intentions as a function of various predictor variables (M0: baseline/no predictors, M1: headline-level ratings, M2: post-level ratings, M3: agreement & alignment interaction, M4: moral framing & user values interaction, M5: moral framing & political ideology interaction). The models had the same structure as shown in Tables 1 - S10.

Additionally, we ran a Bayesian multilevel linear regression model (mediation) to estimate the indirect effects of moral values 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 because analytical thinking could reduce susceptibility to moral framing effects. See Tables 5 for an overview of the model specifications and Table S11 in the SI for the specific R formulas.

Analogous to Study 1, 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 Study 1 (Hypothesis 1), Table 6 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

Table 5
Overview of Bayesian Multilevel Linear Regression Models (Mediation)

Model	Predictors	Coefficients
Deliberation	Headline veracity, familiarity, post's moral framing, participant's moral values, participant's CRT score, deliberation ratings, interactions	Fixed effect of veracity; Fixed and random effects of familiarity, moral framing, moral values, CRT, deliberation, interactions
Response Time	Headline veracity, familiarity, post's moral framing, participant's moral values, participant's CRT score, response time, interactions	Fixed effect of veracity; Fixed and random effects of familiarity, moral framing, moral values, CRT, response time, interactions

Note. Table provides an overview of the specification of the models in this Study. All models include a random intercept for posts, participants, and headlines.

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 ($\beta = 0.11, [0.02, 0.20]; \Delta\beta = 0.14, [0.05, 0.23]$) 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 nonmoral posts (neutral) or posts framed with Individualizing values (misaligned).

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) or posts

Table 6
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 Values	.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*, here, is its Bayesian counterpart.* z *is the standardized difference in out-of-sample prediction accuracy between each model (*$z = \Delta_{ELPD}/SE$*).*

framed 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 S19 for a comparison of effect sizes for Model 4 with and without controls for headline veracity and familiarity. See Table 7 for an overview of the effect sizes of moral alignment vs misalignment and neutral posts on sharing intentions.

Model 4 also predicted sharing intentions more accurately than Model 5 ($\Delta_{ELPD} = 60.22, SE = 16.46, z = 3.66$) which predicted sharing intentions based on political orientation instead of moral values. Overall, Study 2 successfully replicated the facilitatory effect of targeting people's moral values on (mis)information sharing. Consistent with Study 1, 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 Model 4. Across Model 1–3, the most important predictors were, consistent across studies,

Table 7
Effect of participant values on sharing across framing conditions (main model M4)

	Aligned vs Misaligned	Aligned vs Neutral
Binding Values	0.14 [0.05, 0.23]	0.11 [0.02, 0.15]
Individualizing Values	0.15 [0.08, 0.22]	0.13 [0.05, 0.20]

Note. Table shows the difference in the effect of moral alignment vs misalignment or non-moral (neutral) framing on sharing intentions. Table shows a significant effect of moral alignment vs misalignment and neutral posts.

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 1; 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: $\beta = 0.10, [0.07, 0.13]$); agreed with the post (M3: $\beta = 0.17, [0.12, 0.22]$), and considered the post to align with their moral values (M3: $\beta = 0.28, [0.23, 0.33]$).

Lastly, we found no evidence for deliberation mediating the effect of matching posts' moral framing and participant's moral values on sharing intentions (Hypothesis 2). 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 ($\beta = 0.02, [-0.03, 0.08]$; see figure S6 in the SI for a detailed visualization). Furthermore, analytical thinking did not reduce susceptibility to moral framing (Hypothesis 3; $\beta = -0.01, [-0.06, 0.04]; \beta = -0.03, [-0.08, 0.03]$). We also ran an identical mediation analysis with response time for sharing a post as an alternative deliberation measure. We found no effect of matching moral framing and participant 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 ($\beta = 0.01, [-0.03, 0.05]$). For a more detailed analysis of analytical thinking see section 6 in the SI.

Additional Analyses. We conducted additional analyses, replicating the approach used in Study 1b, to investigate whether moral alignment contributes to

explanatory power beyond the headline-level ratings, which were previously identified as strong and consistent predictors of sharing intentions. We again fit an additional model $M1_{total}$ that combines headline-level ratings, moral values, and moral framing. We find that the effect of moral alignment holds up even when controlling for the effects of headline-level ratings. Model $M1_{total}$ predicted sharing intentions more accurately than Model M1 ($\Delta_{ELPD} = 38.69, SE = 14.01, z = 2.76$), which did not include moral alignment, indicating that moral alignment adds explanatory power above and beyond headline-level ratings. Consistent with the main hypothesis, we find again that participants' endorsement of Binding values predicted greater sharing intentions in the Binding framing condition ($\beta = 0.20, [0.12, 0.28]$) than in the Individualizing framing condition ($\beta = 0.10, [0.02, 0.18]; \Delta\beta = 0.10, [0.02, 0.18]$) and in the nonmoral framing condition ($\beta = 0.12, [0.04, 0.20]; \Delta\beta = 0.08, [0.01, 0.16]$). In other words, participants with Binding values had greater sharing intentions for posts framed with Binding values (aligned) than non-moral posts (neutral) and posts with Individualizing values (misaligned). Likewise, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition ($\beta = 0.19, [0.11, 0.26]$) than in the Binding framing condition ($\beta = 0.07, [-0.00, 0.14]; \Delta\beta = 0.12, [0.05, 0.19]$) and in the nonmoral framing condition ($\beta = 0.08, [0.01, 0.15]; \Delta\beta = 0.10, [0.03, 0.17]$). 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).

Note that the previous analyses in this study tested the effects of different predictors on misinformation while only controlling for veracity. In other words, we assessed whether the observed effect of moral alignment is independent of stimuli veracity. However, to make inferences about whether moral alignment specifically facilitates misinformation, we refitted model M4 with a veracity and moral alignment interaction (model $M4_{veracity}$). We then analyzed whether the effects of moral alignment (and misalignment) differed between true and false content to more directly answer the question:

Does targeting audiences' core values facilitate the spread of misinformation?:

Table 8
Effect of participant values on sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.16 [0.07, 0.24]	0.11 [-0.01, 0.23]	0.10 [0.02, 0.19]	0.09, [-0.04, 0.21]
Individualizing Values	0.15 [0.08, 0.22]	0.10 [-0.00, 0.20]	0.12 [0.04, 0.19]	0.10 [-0.01, 0.20]

Note. Table shows the difference in the effect of moral alignment vs misalignment or non-moral (neutral) framing on sharing intentions for both false and true posts. Table shows that the effect of moral alignment is generally larger for false posts than for true posts. Across all conditions the effect of moral alignment on true news is not significant.

We find that the effect of moral alignment differed for true vs false content. That is, for misinformation, participants' endorsement of Binding values predicted greater sharing intentions in the Binding framing condition than in the Individualizing framing condition ($\Delta\beta = 0.16, [0.07, 0.24]$) and in the nonmoral framing condition ($\Delta\beta = 0.10, [0.02, 0.19]$), analogous to the previous findings of model M4. For true information, however, participants' endorsement of Binding values did not predict greater sharing intentions in the Binding framing condition than in the Individualizing framing condition ($\Delta\beta = 0.11, [-0.01, 0.23]$) or in the nonmoral framing condition ($\Delta\beta = 0.09, [-0.04, 0.21]$). In other words, participant showed higher sharing intentions for sharing misinformation (but not true information) framed with Binding values (aligned) than for non-moral posts (neutral) with Individualizing values (misaligned). Likewise, for misinformation, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition than in the Binding framing condition ($\Delta\beta = 0.15, [0.08, 0.22]$) and in the nonmoral framing condition ($\Delta\beta = 0.12, [0.04, 0.19]$). For true information, participants' endorsement of Individualizing values did not predict greater sharing intentions in the Individualizing framing condition than in the Binding

framing condition ($\Delta\beta = 0.10, [-0.00, 0.20]$) or in the nonmoral framing condition ($\Delta\beta = 0.10, [-0.01, 0.20]$). In other words, participants with Individualizing values had greater sharing intentions for misinformation (but not true information) framed with Individualizing values (aligned) than for nonmoral posts (neutral) and posts with Binding values (misaligned). See Table 8 an overview of the effect sizes of moral alignment vs misalignment and neutral posts on sharing intentions.

Summary. Replicating and extending Study 1, Study 2 confirmed that an alignment of a post's moral framing to users' moral values indeed increases 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 whether the post contains true or false content and how familiar the content is. Additional analyses showed that misinformation is more sensitive to moral alignment than true information, for which moral alignment showed no significant effect on sharing intentions (compared to neutral or misaligned posts). Consistent with Study 1, we also 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. Relatedly, trait-level analytical thinking did not moderate the effect of moral alignment, misinformation sharing and plausibility concerns (see our additional analysis in the SI).

Study 3

In Study 3, we analyze COVID-related content on Twitter regarding the relationship between tweets' moral framing, users' political ideology, and liking or sharing of the tweets to test whether the findings from Study 1 and 2 hold in naturalistic online data. We predict that moral framing that matches values associated with a user's political ideology (e.g., liberal and Individualizing values) will lead to increased sharing and liking of

tweets. Since Individualizing values correlate negatively and Binding values correlate positively with political conservatism (see Kivikangas et al., 2021), we expect that content from a conservative source, compared to content from liberal sources, would be shared and liked more frequently when framed with Binding values. Conversely, we expect that content from a liberal source, compared to content from a conservative source, would be shared and liked more frequently 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). Note that whereas Study 3 focuses on the apparent moral values of message sources, Study 1 and Study 2 focused on the moral values of message recipients. However, given that people tend to expose themselves to social media content that agrees with their worldview (Bakshy et al., 2015; González-Bailón et al., 2023) and moral values (Dehghani et al., 2016; Singh et al., 2021), it is very likely that audience engagement measured in Study 3 were captured from users whose moral values matched those of the message source.

Method

We collected social media messages about COVID vaccinations and mandates from Twitter and used natural language processing methods to extract the messages' moral framing. Finally, we fit a model predicting liking and sharing of these messages as a function of messages' moral framing, users' likely political ideology, 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). Using the Twitter IDs provided in this corpus, 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 tweets that contained moral language 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 political ideology using the "Misinformation exposure" API by Mosleh et al. (2021), which returns an ideology score from -1 (liberal) to +1 (conservative) based on political accounts that a user follows.

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[5]:

- Moral Framing: whether the tweet contained moral or non-moral language.
- Binding & Individualizing framing: whether the tweet was framed using Binding and/or Individualizing or non-moral language.
- Political ideology: the tweet source's conservatism on a normalized scale from -1 to 1.

In total, 58% of tweets were posted by conservative users (vs. 42% by liberal users). 28% of tweets contained moral framing (vs. 72% non-moral framing), with 7% of tweets

[5] See Table S2 of the Supplementary Information (SI) for example messages covering the different framing and political ideology.

containing Binding framing and 20% containing Individualizing framing.

Analysis Strategy. 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 its political ideology (e.g., Binding values with conservatives' posts). We ran a series of negative binomial models that predicted the number of retweets or likes as a function of various predictor variables. Model 0 estimated the number of likes as a function of the user's ideology (liberal vs. conservative) and included a fixed intercept and a varying (random) intercept accounting for variance across users. 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 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 users' ideology. We also ran the same series of models with the number of retweets as an alternative outcome variable for user engagement.

We estimated and evaluated these models analogously to Study 1 and Study 2, using the 'brms' `R` package and 10-fold cross-validated ELPD scores.

Results

Table 9 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 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 ideology (liberal and conservative)— predicted engagement more accurately than Model 0 ($\Delta_{ELPD} = 19.75, SE = 6.52, z = 3.03$) and Model 1 ($\Delta_{ELPD} = 19.59, SE = 6.04, z = 3.25$), indicating the relevance of matching moral framing and individuals' values for the spread of social media messages. The between-group analyses demonstrated, as hypothesized, that tweets' Individualizing framing predicted more (1.5 times) engagement when posted by

liberal users compared to conservative users ($\beta^6 = 0.43, [0.24, 0.62]$). However, the difference in engagement for posts with Binding framing when posted by conservative versus liberal users (1.2 times) ($\beta = 0.20, [-0.11, 0.51]$) was not significant. The within-group analyses demonstrated, as hypothesized, that Individualizing framing predicts significantly more engagement (1.8 times) than Binding framing for liberal users ($\beta = 0.56, [0.27, 0.85]$) and there was no difference between both framing for conservative users ($\beta = 0.07, [-0.14, 0.28]$). See an overview of effect sizes and confidence intervals for model M2 in Table 10.

Table 9

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

			z		
Model	Predictors	R^2	M0	M1	M2
M0	Political ideology	0.14	-	-0.03	-3.03
M1	Moral framing, Ideology	0.14	0.03	-	-3.25
M2	Moral framing, Ideology, Interaction	0.14	3.03	3.25	-

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* ($z = \Delta_{ELPD}/SE$).

Analogous to Table 9, Table 11 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 the other models with predictors (M1–M2). Supporting Hypothesis 1, Model 2 —that included tweets' moral framing (Binding and Individualizing) and their interactions with the user's ideology (liberal and conservative)— predicted engagement more accurately than Model 0 ($\Delta_{ELPD} = 33.78, SE = 11.26, z = 3.00$) and Model 1 ($\Delta_{ELPD} = 4.10, SE = 5.80, z = 0.71$). The between-group analyses demonstrated, as hypothesized, that the tweets' Individualizing framing predicted more (2.5 times) engagement when posted by liberal users compared to conservative users

[6] 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 10
Effect of post framing and political ideology on engagement (retweet count)

Between-Group Analysis		Within-Group Analysis	
Framing	Aligned vs Misaligned	Ideology	Aligned vs Misaligned
Binding	0.20 [-0.11, 0.51]	Liberal	0.56 [0.27, 0.85]
Individualizing	0.43 [0.24, 0.62]	Conservative	0.07 [-0.14, 0.28]

Note. The left side of the table shows the difference in the effect of moral alignment vs misalignment on sharing intentions for posts with Binding and Individualizing framing (Between-Group Analysis). Posts with Binding framing are aligned with conservative (rather than liberal) users, while posts with Individualizing framing are aligned with liberal (rather than conservative) users. The right side of the table shows the difference in the effect of moral alignment vs misalignment on sharing intentions for conservative and liberal users message framing (Within-Group Analysis). For liberal users, Individualizing framing is aligned, and for conservative users, there should be no prioritization of one value/framing over the other.

($\beta = 0.89, [0.42, 1.37]$). However, the difference in engagement between conservative and liberal users for posts with Binding framing (1.3 times) ($\beta = 0.27, [-0.46, 0.99]$) was not significant. The within-group analyses demonstrated, as hypothesized, that Individualizing framing predicted significantly more engagement (1.5 times) compared to Binding framing for liberal users ($\beta = 0.41, [0.21, 0.60]$). We also found that conservative users received more engagement (2.10 times) for posts with Binding compared to Individualizing framing ($\beta = 0.74, [0.20, 1.29]$). See an overview of the effect sizes and confidence intervals for model M2 in Table 12.

Overall, these findings show that an alignment of moral framing and political ideology, within and across groups (liberals and conservatives), increases engagement with social media messages. For example, Individualizing framing facilitated engagement for liberals' (compared to conservatives') tweets, and liberals' tweets with individualizing framing received higher engagement than with Binding framing. For Binding framing, the results were somewhat less pronounced. While the effects where in the expected direction, the differences between liberals and conservatives were not statistically significant. However, as expected there was no prioritization of either framing for conservatives. The

Table 11
Comparison of models estimating engagement (favourites count) as a function of various predictor variables

			z		
Model	Predictors	R^2	M0	M1	M2
M0	Political ideology	0.13	-	-2.82	-3.00
M1	Moral framing, Ideology	0.13	2.82	-	-0.71
M2	Moral framing, Ideology, Interaction	0.14	3.00	0.71	-

Note. R^2*, here, is its Bayesian counterpart.* z *is the standardized difference in prediction accuracy between each model (*$z = \Delta_{ELPD}/SE$*).*

lack of difference for Binding framing might be due to conservatives endorsing both Individualizing and Binding values (albeit less strongly than liberals for Individualizing values) thus having no clearly "misaligned" condition the way liberals do. Additionally in Study 3, unlike Study 1 and Study 2 in which we re-framed shared headlines to keep the underlying information constant, we were not able to separate the arguments made in the respective tweets from their framing. For example, "pro-vax" tweets shared with Binding framing might still elicit engagement from liberals but not from conservatives because the underlying pro-vax argument was more strongly associated with liberals than conservatives.. Lastly, there might be differences in public and private sharing that led to different result patterns for the conservative within-group analyses as retweets show up publicly on ones feed whereas information about ones liked tweets is less publicly available and thus more aligned with private sharing. We investigated this in our supplemental analysis (see Section 8 in the SI), in which we repeated our analysis of Study 1 and Study 2 for public and private sharing intentions separately. We find that, generally, the effect of moral alignment is most pronounced for public online sharing, followed by private online sharing, and barely present for private offline sharing (see the detailed overview of models and coefficients in the SI). Our supplemental findings emphasize the social underpinnings of sharing intentions and their connection to aligning with social motivations as additional contributors to sharing intentions. In other words, individuals may not only agree more

Table 12
Effect of post framing and political ideology on engagement (favorite count)

Between-Group Analysis		Within-Group Analysis	
Framing	Aligned vs Misaligned	Ideology	Aligned vs Misaligned
Binding	0.27 [-0.46, 0.99]	Liberal	0.41 [0.21, 0.60]
Individualizing	0.89 [0.42, 1.37]	Conservative	0.74 [0.20, 1.29]

Note. The left side of the table shows the difference in the effect of moral alignment vs misalignment on sharing intentions for posts with Binding and Individualizing framing (Between-Group Analysis). Posts with Binding framing are aligned with conservative (rather than liberal) users, while posts with Individualizing framing are aligned with liberal (rather than conservative) users. The right side of the table shows the difference in the effect of moral alignment vs misalignment on sharing intentions for conservative and liberal users message framing (Within-Group Analysis). For liberal users, Individualizing framing is aligned, and for conservative users, there should be no prioritization of one value/framing over the other.

with morally matched content but may also have a desire to express it to others. This aligns with previous research, including work by C. S. Lee and Ma (2012) and Wong and Burkell (2017), which incorporates social determinants of news sharing, such as status seeking, self-expression, and expressing group ties.

General Discussion

Across three studies, two behavioral experiments and one large-scale analysis of real-world conversations on Twitter, we found that a match of framing and values led to increased sharing of (mis)information. Crucially, these effects were found while controlling for message veracity and familiarity.

Our findings indicate that it is not just moral content but rather *matched* moral content that matters. Importantly, our experimental manipulation 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 should not charge for evacuations, the negative sentiment, or the left-leaning viewpoint. We created our stimuli through carefully crafting matched messages while staying away from obviously partisan headlines 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 any role in (mis)information sharing. Instead, our findings simply demonstrate that moral values can affect (mis)information sharing independent of political ideology. Additionally, moral values could amplify the effects of ideology. For example, in the real world, partisan messages are frequently and differentially accompanied by moralized language and arguments (Fulgoni et al., 2016; Mokhberian et al., 2020), which can then contribute to and amplify partisan differences in misinformation sharing as observed by prior research (Kaplan et al., 2021; Van Bavel & Pereira, 2018; Winkielman et al., 2012). Note, however, that although the connection between moral values and political ideology has been firmly established, there are multiple perspectives on the direction of this relationship. Some research suggests a reversed relationship, with political ideology explaining moral values (Hatemi et al., 2019; Strupp-Levitsky et al., 2020).

Past work on misinformation has shown that more deliberative, analytical reasoning leads to less sharing of misinformation, indicating that more analytical individuals might be able override initial sharing intentions (Bronstein et al., 2019; Mosleh et al., 2021; Pennycook & Rand, 2019a, 2019b). However, in Study 2, analytical and "lazy" thinkers did not differ in their sharing of misinformation and how much they relied on plausibility cues. Furthermore, deliberation 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 and deliberation. It is possible that the effectiveness of analytical thinking is

restricted to contexts that do not strongly evoke values, group identities, and threats thereof (e.g., see S. Lee et al. (2020), Osmundsen et al. (2021), Pretus et al. (2022), and Tandoc et al. (2021) for failures to replicate the effect of analytical thinking and to exclude motivational drivers). 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 6 of the SI found an effect of analytical thinking on misinformation sharing 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 the question of 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 perceived moral alignment with a post may be a motivational driver behind misinformation sharing, potentially further amplified by moral-emotional responses to aligned moral framing of posts. Some evidence for this idea comes from past research that found a facilitatory link between emotional responses and a lack of analytical thinking on believing and sharing of misinformation (Li et al., 2022; Wang et al., 2020). Deliberation might be used, only for strong enough accuracy concerns or weak value and identity-based motives, to "rethink" 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 the processing, believing, and sharing of misinformation (Van Damme & Smets, 2014). Our work confirms past findings that moral content elicits more engagement on social media platforms compared to non-moral 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 in the field of misinformation.

However, our work comes with some limitations. Although our stimuli are arguably naturalistic – that is, we analyzed real-world Twitter data in Study 3 and used realistic posts (including real news headlines) in Study 1 and Study 2 – 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 1 and Study 2 with no other 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; González-Bailón et al., 2023). 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 see (e.g., voting patterns or health-related behaviors).

Additionally, we did not analyze the effect of (perceived) group identity as a potential additional mediator for the effect of moral alignment on misinformation sharing. Social Identity theory suggests that people might implicitly trust a post by someone who is in their ingroup (Hogg, 2016; Tanis & Postmes, 2005), and past work has shown that this can also apply to more misinformation sharing (Ecker et al., 2022; Mackie et al., 1990). Simultaneously, people might perceive someone as an ingroup member when expressing similar values (i.e., via moral framing). Specifically, on social media, users usually do not directly know the posters' actual ideology and values, but instead, infer it from their message or other public features (e.g., profile picture, name). Our work demonstrates that moral alignment leads to increased agreement with a post and, consequently, sharing intentions, which may be linked to perceived group identities. However, this study primarily focuses on the more direct effects of moral alignment on sharing intentions, leaving analyses of perceived group identity for future work.

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 behaviors are monetized and critical to their financial models (Anderson & Wood, 2021; Bayer et al., 2022; Docherty, 2020). Social rewards (e.g., likes) 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 but is not necessarily accurate. This results in a significant proportion of misinformation online being shared by highly habitual users (Ceylan et al., 2023). Future work should investigate the role that moral values and message content play 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. For example, Pennycook and Rand (2021) found that users' sharing intentions of false headlines were significantly higher than their accuracy ratings, potentially indicating habitual sharing of headlines independent of accuracy judgments. In this way, cognitive factors, socio-affective factors and habits might tie into an integrated system of sharing and believing misinformation online.

We hope that our work can facilitate the development of effective countering mechanisms to combat misinformation spread, similar to other harmful messages, such as hate speech and conspiracy theories (Cinelli et al., 2022; Windisch et al., 2022). Most of these countermeasures will have to be implemented at the platform level. For example, platforms could incorporate our results when designing algorithms to curate user feeds by dampening the extent to which content that contains highly emotional and moralized language, detected via natural language processing, is promoted in order to make campaigns based on framing and specific language less effective. Future work could also test the efficacy of inoculation against misinformation, incentivizing the sharing of truthful information, or similar strategies against moral alignment to improve countermeasures at the user level.

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 facilitate the spread of misinformation. 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 the 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.

Constraints on Generality

We acknowledge limitations in the generalizability of our findings regarding the spread of misinformation through moral framing on social media. Study 3 analyzed Twitter users' conversations about COVID vaccinations, with an identified distribution of 64% anti-vaccination and 36% pro-vaccination tweets posted by 58% identified conservative and 42% liberal users, but lacked detailed personal demographics about the users who liked and retweeted those tweets. The data was sourced from a pre-existing, academically reviewed COVID misinformation and rumors dataset, limiting our insights to the types of

misinformation previously identified within this data. Studies 1 and 2 involved participants balanced for gender (male, female) and political orientation (liberal, moderate, conservative), yet may not fully represent the broader population's diversity in other demographics. Therefore, while our findings offer significant insights into the interplay of moral framing and misinformation spread, they are primarily applicable to social media users engaged in sharing and consuming news and related content, and caution is advised in extending these findings to other contexts or demographics.

Author Contributions

S.A. lead conceptualization, data curation, software, formal analysis, investigation, methodology, visualization, writing the original draft, and reviewing & editing the manuscript. N.R. lead resources, contributed equally to conceptualization, data curation, software, formal analysis, investigation, methodology, visualization, and supported reviewing & editing the manuscript. P.G. supported conceptualization, data curation, reviewing & editing the manuscript, and contributed equally to software. E.B. supported conceptualization, methodology, resources, and reviewing & editing the manuscript. Y.S. supported conceptualization, resources, and reviewing & editing the manuscript. J.T. supported conceptualization, resources, and reviewing & editing the manuscript. R.L. supported conceptualization, resources, and reviewing & editing the manuscript. J.K. supported conceptualization, methodology, resources, and reviewing & editing the manuscript. C.P. supported conceptualization, methodology, resources, and reviewing & editing the manuscript. M.D. contributed equally to conceptualization, supported methodology, resources, reviewing & editing the manuscript, and lead funding acquisition, supervision, and project administration.

References

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

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

Chollet, F. et al. (2015). *Keras*. https://github.com/fchollet/keras

Ciampaglia, G. L., Mantzarlis, A., Maus, G., & Menczer, F. (2018). Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine*, *39*(1), 65–74.

Cinelli, M., Etta, G., Avalle, M., Quattrociocchi, A., Di Marco, N., Valensise, C., Galeazzi, A., & Quattrociocchi, W. (2022). Conspiracy theories and social media platforms. *Current Opinion in Psychology*, 101407.

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.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature human behaviour*, *1*(11), 769–771.

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.

Dong, Z. S., Meng, L., Christenson, L., & Fulton, L. (2021). Social media information sharing for natural disaster response. *Natural hazards*, *107*, 2077–2104.

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.

Elsey, J. (2021). Powerful sequential designs using bayesian estimation: A power analysis tutorial using brms, the tidyverse, and furrr.

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.

Erickson, L. B. (2011). Social media, social capital, and seniors: The impact of facebook on bonding and bridging social capital of individuals over 65.

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.

Fulgoni, D., Carpenter, J., Ungar, L., & Preoiuc-Pietro, D. (2016). An empirical exploration of moral foundations theory in partisan news sources. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3730–3736.

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

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., et al. (2023). Asymmetric ideological segregation in exposure to political news on facebook. *Science*, *381*(6656), 392–398.

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.

Hatemi, P. K., Crabtree, C., & Smith, K. B. (2019). Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, *63*(4), 788–806.

Hochschild, J. L., & Einstein, K. L. (2015). Do facts matter? information and misinformation in american politics. *Political Science Quarterly*, *130*(4), 585–624.

Hogg, M. A. (2016). *Social identity theory*. Springer.

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.

Joyal-Desmarais, K., Scharmer, A. K., Madzelan, M. K., See, J. V., Rothman, A. J., & Snyder, M. (2022). Appealing to motivation to change attitudes, intentions, and behavior: A systematic review and meta-analysis of 702 experimental tests of the effects of motivational message matching on persuasion. *Psychological Bulletin*, *148*(7-8), 465.

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.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in human behavior*, *28*(2), 331–339.

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.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, ... Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org]. https://www.tensorflow.org/

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020). Moral framing and ideological bias of news. *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, 206–219.

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on twitter. *Nature communications*, *12*(1), 921.

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.

Oh, S., & Syn, S. Y. (2015). Motivations for sharing information and social support in social media: A comparative analysis of f acebook, t witter, d elicious, y ou t ube, and f lickr. *Journal of the Association for Information Science and Technology*, *66*(10), 2045–2060.

Osatuyi, B. (2013). Information sharing on social media sites. *Computers in Human Behavior*, *29*(6), 2622–2631.

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., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., & Rand, D. G. (2019a). 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. (2019b). 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.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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.

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative ai. *PNAS Nexus*, pgae035.

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.

Strupp-Levitsky, M., Noorbaloochi, S., Shipley, A., & Jost, J. T. (2020). Moral foundations as the product of motivated social cognition: Empathy and other psychological

underpinnings of ideological divergence in individualizing and binding concerns. *PloS one*, *15*(11), e0241144.

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.

Tanis, M., & Postmes, T. (2005). A social identity approach to trust: Interpersonal perception, group membership and trusting behaviour. *European journal of social psychology*, *35*(3), 413–424.

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., & Ramírez, 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.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

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.

Windisch, S., Wiedlitzka, S., Olaghere, A., & Jenaway, E. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell systematic reviews*, *18*(2), e1243.

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.

Wong, L. Y., & Burkell, J. (2017). Motivations for sharing news on social media. *Proceedings of the 8th International conference on social media & society*, 1–5.

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.

Supplemental Material for

Targeting Audiences' Moral Values Shapes Misinformation Sharing

1 Stimuli

Study 1 & Study 2

Table S1
List of headlines selected for Study 1 and Study 2

#	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 1a.

Study 3

Table S2

Exemplary tweets showcasing moral framing and/or political ideology on COVID vaccinations and mandates

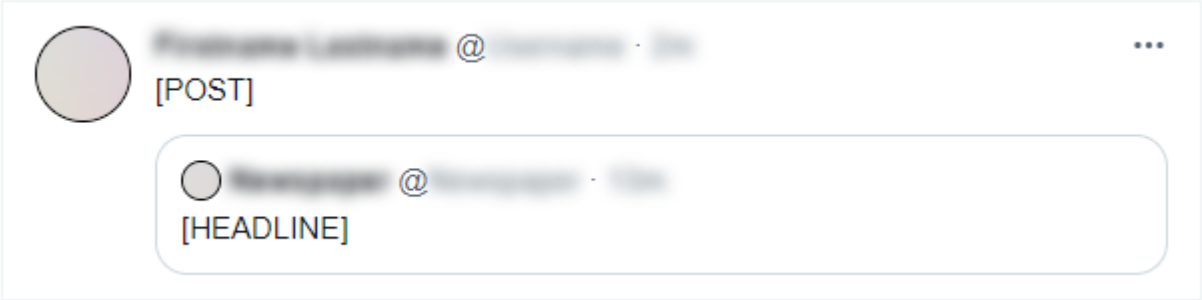
Topic	Description	Example
Nonmoral	Does not contain moral language	They really think mandating vaccines on airplanes is gonna sway the unvaccinated, lol. I guess Im gonna just drive...
Individualizing	Focused on individual rights and well-being	Under 12s are unvaccinated! We need to ensure all primary 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!
Liberal user (pro-vax)	Endorsing COVID vaccines and mandates	Lets get vaxxed!
Conservative user (anti-vax)	Opposing COVID vaccines and mandates	No. Do Not get Vaccinated!

2 Questionnaire Items (Study 1a, 1b, 2)

Introduction

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

Figure S1
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 S2
Example stimulus presentation for headline-level items

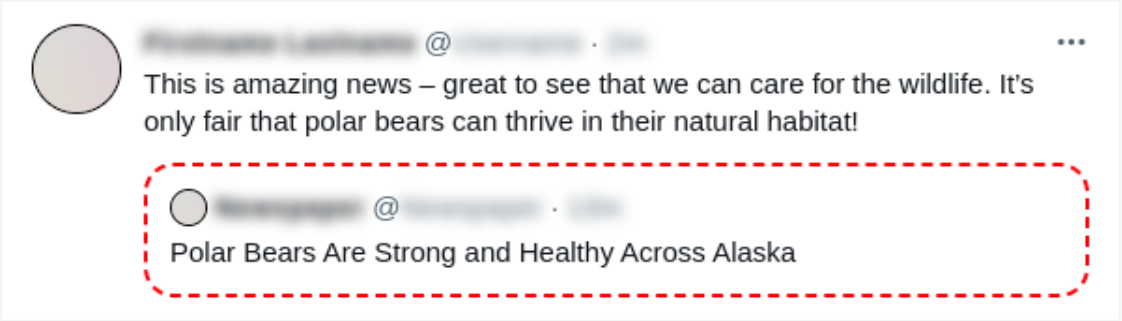


Table S3
In your opinion, the post the user has written about the headline is

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 S3
Example stimulus presentation for post-level items

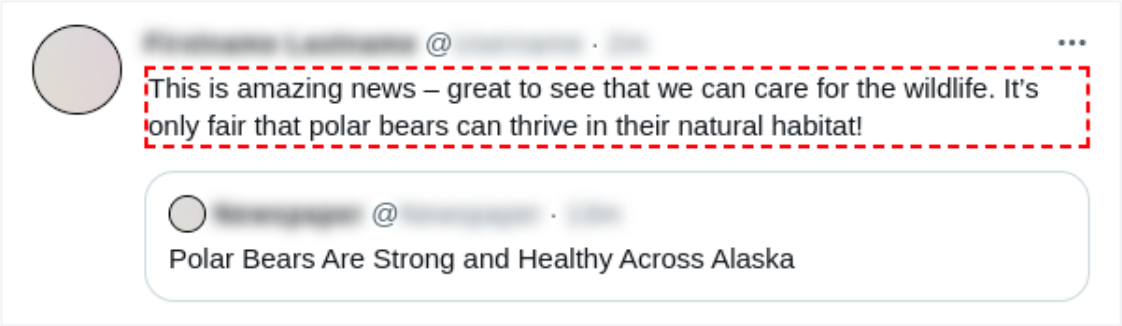


Table S4
In your opinion, the post the user has written about the headline is

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 S5
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 S6
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 S4
Example stimulus presentation for whole-stimulus items

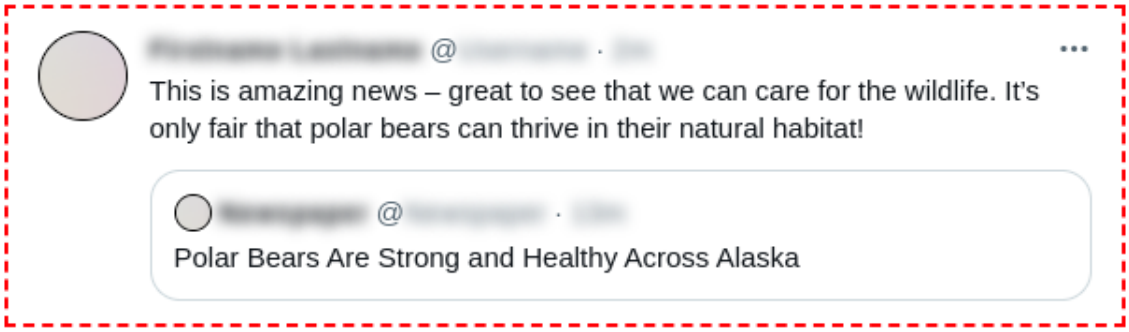


Table S7
If you came across this post, how likely would you be to ...

... 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 S8
Why did you decide to share (or not to share) the previous post? (Only Study 2)

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

3 Study 3 Robustness Checks

The data in Study 3 was sampled from was a corpus originally collected in an effort to study misinformation, rumors, and conspiracy theories regarding COVID vaccinations and mandates. However, this data was collected using keywords, which might have led to a significant number of truthful content being included in this data. As a robustness check, we determined the average level of exposure to misinformation for users in our dataset as a proxy for the prevalence of misinformation in our data. We used the API previously used to retrieve users' political ideology (Mosleh et al., 2021) to retrieve users' exposure score from 0 (no exposure) to 1 (complete exposure), based on the accounts the user follows, the credibility of sources that these accounts share, as well as the amount of previously identified misinformation shared by these accounts. We find an average misinformation exposure score of $M = 0.59$. As a comparison, Mosleh et al. (2021) identified clusters of low-exposed users with an average exposure score of $M = 0.389$ and a highly-exposed cluster with an average exposure score of $M = 0.501$, indicating that our data likely has a high prevalence of misinformation.

4 Detailed Model Formulas

Study 1b

Table S9

R Formulas for Bayesian Multilevel Linear Regression Models

Model	R Formula
M0	`brm(z_post_share_index  1 + (1|ii) + (1|jj) + (1|kk))`
M1	`bbrm(z_post_share_index 1 + z_kk_headline_believable + z_kk_headline_controversial + z_kk_headline_surprising + z_kk_headline_interesting + z_kk_headline_positive + z_kk_headline_true + z_headline_believable + z_headline_controversial + z_headline_surprising + z_headline_interesting + z_headline_positive + (1|ii) + (1|jj) + (1 + z_headline_believable + z_headline_controversial + z_headline_surprising + z_headline_interesting + z_headline_positive|kk))`
M2	`brm(z_post_share_index 1 + z_ii_post_controversial + z_ii_post_surprising + z_ii_post_interesting + z_ii_post_positive + z_post_controversial + z_post_surprising + z_post_interesting + z_post_positive + (1 + z_post_controversial + z_post_surprising + z_post_interesting + z_post_positive|ii) + (1|jj) + (1|kk))`
M3	`brm(z_post_share_index  1 + z_post_agree*z_post_align + z_ii_post_agree*z_ii_post_align + (1 + z_post_agree*z_post_align|ii) + (1|jj) + (1|kk))`
M4	`brm(z_post_share_index  1 + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_post_framing*z_jj_mfq_prop + (1|jj) + (1 + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_post_framing*z_jj_mfq_prop|kk))`
M5	`brm(z_post_share_index  1 + x_post_framing*z_jj_conservatism + (1|jj) + (1 + x_post_framing*z_jj_conservatism|kk))`

Note. Table provides an overview of the R formulas of the models in this Study. "kk" indicates headline-level, "ii" indicates post-level, and "jj" indicates a user-level variable. See the R code for details on the implementation, such as number of chains, warm-up, etc.

Table S10

R Formulas for Bayesian Multilevel Linear Regression Models (Mediation)

Model	R Formula
Mediation	`brm(bf(z_post_agree 1 + z_headline_true + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + (1|jj) + (1 + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind|kk)) + bf(z_post_align 1 + z_headline_true + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + (1|jj) + (1 + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind|kk)) + bf(z_post_share_index 1 + z_headline_true + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_post_framing*z_post_agree*z_post_align + (1|jj) + (1 + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_post_framing*z_post_agree*z_post_align|kk)))`

Note. Table provides an overview of the R formula of the mediation model in this Study. "kk" indicates headline-level, "ii" indicates post-level, and "jj" indicates a user-level variable. See the R code for details on the implementation, such as number of chains, warm-up, etc.

Study 2

Table S11

R Formulas for Bayesian Multilevel Linear Regression Models (Mediation)

Model	R Formula
Deliberation	`brm(bf(z_post_deliberation_index 1 + x_headline_true + z_headline_familiar + z_jj_crt*x_post_framing*z_jj_mfq_indi + z_jj_crt*x_post_framing*z_jj_mfq_bind + (1|jj) + (1 + z_jj_crt*x_post_framing*z_jj_mfq_indi + z_jj_crt*x_post_framing*z_jj_mfq_bind + z_headline_familiar|kk)) + bf(z_post_share_index 1 + z_headline_familiar + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_headline_true*x_post_framing*z_post_deli + (1|jj) + (1 + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_post_framing*z_post_deliberation_index + z_headline_familiar|kk)))`
Response Time	`brm(bf(z_post_response_time 1 + x_headline_true + z_headline_familiar + z_jj_crt*x_post_framing*z_jj_mfq_indi + z_jj_crt*x_post_framing*z_jj_mfq_bind + (1|jj) + (1 + z_jj_crt*x_post_framing*z_jj_mfq_indi + z_jj_crt*x_post_framing*z_jj_mfq_bind + z_headline_familiar|kk)) + bf(z_post_share_index 1 + z_headline_familiar + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_headline_true*x_post_framing*z_post_resp + (1|jj) + (1 + x_post_framing*z_jj_mfq_indi + x_post_framing*z_jj_mfq_bind + x_post_framing*z_post_response_time + z_headline_familiar|kk)))`

Note. Table provides an overview of the R formula of the mediation models in this Study. "kk" indicates headline-level, "ii" indicates post-level, and "jj" indicates a user-level variable. See the R code for details on the implementation, such as number of chains, warm-up, etc.

5 Additional Analyses of Inferring Political Ideology via Stance on Vaccination

In Study 3 we determined users' political ideology via their past retweeting behavior and the accounts they follow (i.e., based on what verified political accounts a user follows and shares). However, this approach might be limited because it requires to have knowledge about the ideology about individuals in a user's network at a given time and mainly relies on a number of known political accounts (usually famous politicians, pundits, or organizations). Instead one could also determine a users ideology based on the contents they share specific to the conversation at hand, here content that is "anti-vax" or "pro-vax". The benefit of this approach is that users' ideology can be determined directly from their behavior in the course of relevant posts, such as posts about vaccinations, and during the same time frame that the data is being collected. Of course, this in turn relies on the topic being sufficiently politically polarized, i.e., "anti-vax" being strongly correlated with conservatism and "pro-vax" being strongly correlated with being liberal (Clarkson & Jasper, 2022; Jiang et al., 2021; Kerr et al., 2021; Stroope et al., 2021).

In an robustness check to the analysis to Study 3, we infer ideology via users' stance on vaccination and test whether ideology measured via the misinformation-exposure API explains as much or more variance compared to our proxy. We 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 refit the models from Study 3 using stance instead of political ideology and fit a model that adds stance to the ideology x moral values model, and compare whether ideology explains as much or more

variance compared to stance and whether stance has any explanatory power beyond ideology.

We find that a model using political ideology has significantly higher explanatory power than a model using stance ($\Delta_{ELPD} = 20.06, SE = 5.85, z = 3.43$) and stance did not add any explanatory power to the ideology x moral values model ($\Delta_{ELPD} = 22.31, SE = 14.08, z = 1.58$), indicating that stance was indeed simply a proxy for ideology. Furthermore, this robustness check adds additional credibility to the political ideology score determined by the misinformation-exposure API by showing alignment with alternative measures of ideology.

6 Additional Analyses of Analytical Thinking

Study 2 replicated the results of Study 1 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

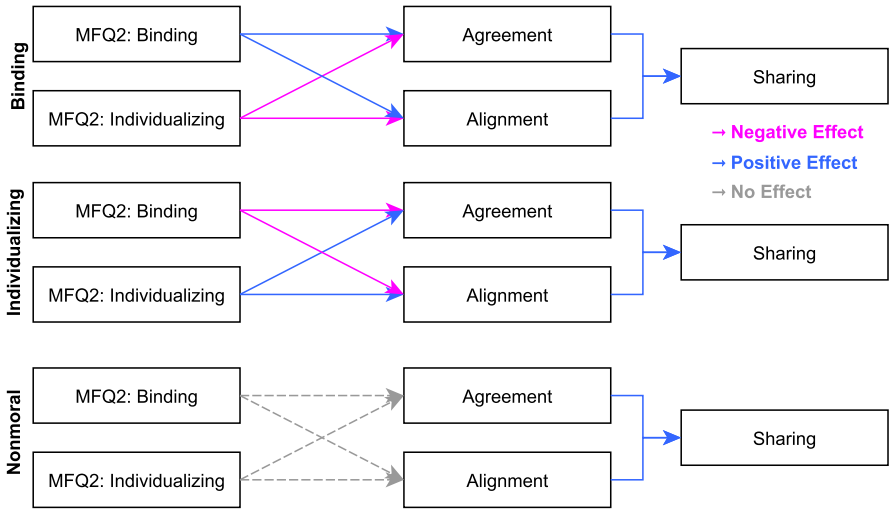
($\Delta_{ELPD} = -2.61, SE = 8.52, z = -0.31; \Delta_{ELPD} = 7.70, SE = 9.04, z = 0.85$), which predicted sharing intentions as a function of random intercepts for participants, headlines, and posts. Third, we fitted model M9 that predicted sharing intentions as a function of headline veracity, analytical 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, 2019b). 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, 2019b). 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 1, 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 1, 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 S5 provides an overview of the observed relationships. Supporting the original Hypothesis 2 in Study 1, 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 S5

Replication of the mediation in Study 1: Effect of matching moral values and framing via perceived agreement and alignment with the post



Note. The figure shows a clear separation between the effect of matching moral framing, which leads to an increase of sharing intentions (blue color), mismatched moral framing, which leads to a decrease of sharing intentions (red color), and not addressing moral values, which has no effect on sharing intentions (grey color), across all conditions.

7 Additional Analyses of Stimuli Order Effects

In our studies 1 & 2, 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: $\beta_{order:framing1:individualizing} = 0.00, [-0.02, 0.02]$, $\beta_{order:framing2:binding} = 0.01, [-0.02, 0.03]$).

8 Additional Analyses of Public vs Private Sharing

In our studies 1 & 2, we utilized an index that combined public and private sharing intentions (public sharing, public liking, private online sharing, private offline sharing). While this index had high reliability above 0.8, the underlying items might tap into distinct motivations of sharing. Given people's motivation to express their values and other potential social functions of public sharing, such as aligning with group values and identities (Oh & Syn, 2015)it could be that moral alignment is primarily drives public sharing intentions and not private and offline sharing which do not fulfill the same sharing motivations (e.g., difficult to signal to the group through private or offline sharing).

Thus, we refit the main model $M4_{veracity}$ (which includes moral alignment and its interaction with veracity) from Study 1 and Study 2 respectively using each index as a separate outcome variable and analyze whether the observed effects of moral alignment drives either hold for public, private sharing, or both.

Study 1

We find that the previously reported effects for moral alignment only hold up for public sharing. In this case, model $M4_{veracity}$ predicted sharing intentions significantly more accurate compared to the baseline model M0 ($\Delta_{ELPD} = 66.64, SE = 17.32, z = 3.85$) and model M5 ($\Delta_{ELPD} = 33.67, SE = 17.18, z = 1.96$) which uses political ideology instead of moral values. Consistent with the results in Study 1, we find that participants Binding values predicted significantly more public sharing intentions for false posts framed with Binding compared to Individualizing values $\Delta\beta = 0.19, [0.10, 0.29]$) and, to a lesser extent, non-moral posts $\Delta\beta = 0.09, [-0.01, 0.18]$). For true information however, participants' endorsement of Binding values did not predict greater sharing intentions in the Binding framing condition than in the Individualizing framing condition ($\Delta\beta = 0.09, [-0.05, 0.23]$) or in the nonmoral framing condition ($\Delta\beta = 0.09, [-0.05, 0.22]$). In other words, participant showed higher sharing intentions for sharing misinformation (but not true information) framed with Binding values (aligned) than posts with Individualizing values (misaligned). Likewise, participants' endorsement of Individualizing values predicted greater sharing intentions, for false posts, in the Individualizing framing condition than in the Binding framing condition ($\Delta\beta = 0.20, [0.11, 0.28]$) and in the nonmoral framing condition ($\Delta\beta = 0.11, [0.01, 0.21]$). For true information, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition than in the Binding framing condition ($\Delta\beta = 0.18, [0.06, 0.30]$) and, to a lesser extent, in the nonmoral framing condition ($\Delta\beta = 0.09, [-0.05, 0.23]$). In other words, participants with Individualizing values had greater sharing intentions for misinformation framed with Individualizing values (aligned) than nonmoral posts (neutral) and posts with Binding values (misaligned) but for true information this effect was dampened and participants had only greater sharing intentions for posts framed with Individualizing values (aligned) vs Binding (misaligned). Note again that the effect sizes of moral alignment were, across all conditions, lower for true information compared to

misinformation even when the effects were still significant (e.g., Individualizing vs Binding framing for participants with Individualizing values), indicating a generally lower sensitivity of true information to moral alignment.

Table S12

Effect of participant values on public online sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.19 [0.10, 0.29]	0.09 [-0.05, 0.23]	0.09 [-0.01, 0.18]	0.09 [-0.05, 0.22]
Individualizing Values	0.20 [0.11, 0.28]	0.18 [0.06, 0.30]	0.11 [0.01, 0.21]	0.09 [-0.05, 0.23]

Note. Table shows the difference in effect of moral alignment vs misalignment or non-moral (neutral) framing on public sharing intentions for both false and true posts. Table shows that the effect of moral alignment is generally larger for false posts than for true posts and that the effect of moral alignment is generally not significant for true posts, indicating that misinformation is more sensitive to moral alignment.

For private online and offline sharing, model $M4_{veracity}$ did not predict sharing intentions more accurately than the baseline model M0 ($z_{private} = 1.28; z_{offline} = -1.64$) or model M5 ($z_{private} = 1.42; z_{offline} = -2.12$) which uses political ideology instead of moral values. Furthermore, we did not find an effect of moral alignment on sharing of either true or false posts, except for aligned Individualizing values vs non-moral and misaligned posts. See Table S13 and Table S14 for an overview of the effects across all conditions and stimuli veracity for private online and offline sharing respectively.

Study 2

We find that the previously reported effects for moral alignment only hold up for public sharing. In this case, model $M4_{veracity}$ predicted sharing intentions significantly more accurate compared to the baseline model M0 ($\Delta_{ELPD} = 63.66, SE = 16.74, z = 3.80$) and model M5 ($\Delta_{ELPD} = 48.91, SE = 16.91, z = 2.89$) which uses political ideology instead of moral values. Consistent with the results in Study 1, we find that participants Binding

Table S13

Effect of participant values on private online sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.02 [-0.07, 0.11]	0.03 [-0.11, 0.16]	0.00 [-0.10, 0.10]	0.03, [-0.11, 0.16]
Individualizing Values	0.11 [0.03, 0.18]	0.09 [-0.02, 0.20]	0.10 [0.02, 0.18]	0.07 [-0.04, 0.19]

Note. Table shows the difference in effect of moral alignment vs misalignment or non-moral (neutral) framing on public sharing intentions for both false and true posts. Table shows no effect of moral alignment except for Individualizing values on Individualizing vs Binding and non-moral framing for false posts. Results indicate that private sharing is not driven by moral alignment.

Table S14

Effect of participant values on private offline sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.03 [-0.07, 0.12]	0.02 [-0.11, 0.15]	0.02 [-0.08, 0.11]	0.08 [-0.05, 0.21]
Individualizing Values	0.07 [-0.00, 0.15]	0.02 [-0.09, 0.13]	0.01 [-0.07, 0.09]	-0.04 [-0.16, 0.08]

Note. Table shows the difference in effect of moral alignment vs misalignment or non-moral (neutral) framing on public sharing intentions for both false and true posts. Table shows no effect of moral alignment except across all conditions and for both true and false posts. Results indicate that private offline sharing is not driven by moral alignment.

values predicted significantly more public sharing intentions for false posts framed with Binding compared to Individualizing values $\Delta\beta = 0.18, [0.09, 0.26]$) and non-moral posts $\Delta\beta = 0.12, [0.03, 0.21]$). For true information however, participants' endorsement of Binding values did not predict greater sharing intentions in the Binding framing condition than in the Individualizing framing condition ($\Delta\beta = 0.12[-0.00, 0.24]$) or in the nonmoral framing condition ($\Delta\beta = 0.11[-0.02, 0.24]$). In other words, participant showed higher sharing intentions for sharing misinformation (but not true information) framed with Binding values (aligned) than non-moral posts (neutral) and posts with Individualizing values (misaligned). Likewise, participants' endorsement of Individualizing values predicted greater sharing intentions, for false posts, in the Individualizing framing condition than in the Binding framing condition ($\Delta\beta = 0.19[0.11, 0.26]$) and in the nonmoral framing condition ($\Delta\beta = 0.14, [0.06, 0.22]$). For true information, participants' endorsement of Individualizing values predicted greater sharing intentions in the Individualizing framing condition than in the Binding framing condition ($\Delta\beta = 0.12, [0.01, 0.22]$) and, albeit not significantly, in the nonmoral framing condition ($\Delta\beta = 0.12[-0.00, 0.23]$). In other words, participants with Individualizing values had greater sharing intentions for misinformation framed with Individualizing values (aligned) than nonmoral posts (neutral) and posts with Binding values (misaligned) but for true information this effect was dampened and participants had only greater sharing intentions for posts framed with Individualizing values (aligned) vs Binding (misaligned). Note again that the effect sizes of moral alignment were, across all conditions, lower for true information compared to misinformation even when the effects were still significant (e.g., Individualizing vs Binding framing for participants with Individualizing values), indicating a generally lower sensitivity of true information to moral alignment.

For private online and offline sharing, model $M4_{veracity}$ did not predict sharing intentions more accurately than the baseline model M0 ($z_{private} = 0.35; z_{offline} = 0.25$) or model M5 ($z_{private} = 0.54; z_{offline} = 0.54$) which uses political ideology instead of moral

Table S15

Effect of participant values on public online sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.18 [0.09, 0.26]	0.12 [-0.00, 0.24]	0.12 [0.03, 0.21]	0.11 [-0.02, 0.24]
Individualizing Values	0.19 [0.11, 0.26]	0.12 [0.01, 0.22]	0.14 [0.06, 0.21]	0.12 [-0.00, 0.23]

Note. Table shows the difference in effect of moral alignment vs misalignment or non-moral (neutral) framing on public sharing intentions for both false and true posts. Table shows that the effect of moral alignment is generally larger for false posts than for true posts and that the effect of moral alignment is only significant for Individualizing values and framing.

values. In line with the results for Study 1, we found no effect of moral alignment on private offline sharing, supporting our previous conclusion about separate underlying mechanisms for this kind of sharing. Similar to Study 1, we find an effect of moral alignment (vs misalignment) for private online sharing of misinformation but not true information similar to the results of public sharing. However, the effect sizes are smaller compared to those for public sharing and there is no effect for moral alignment compared to non-moral (neutral) posts. These findings indicate that there might distinct underlying dynamics driving public online sharing, private online sharing, and private offline sharing, with moral alignment mainly facilitating public online sharing and partially facilitating private online sharing but not offline sharing. See Table S16 and Table S17 for an overview of the effects across all conditions and stimuli veracity for private online and offline sharing respectively.

Table S16

Effect of participant values on private online sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.10 [0.01, 0.19]	0.06 [-0.07, 0.19]	0.06 [-0.03, 0.16]	0.05, [-0.08, 0.18]
Individualizing Values	0.14 [0.07, 0.22]	0.10 [-0.00, 0.20]	0.10 [0.02, 0.17]	0.08 [-0.03, 0.19]

Note. Table shows the difference in effect of moral alignment vs misalignment or non-moral (neutral) framing on public sharing intentions for both false and true posts. Table shows an effect of moral alignment vs misalignment for false but not true posts. Results indicate that moral alignment might facilitate private online sharing of misinformation.

Table S17

Effect of participant values on private offline sharing across framing conditions and stimuli veracity

	Aligned vs Misaligned		Aligned vs Neutral	
	False	True	False	True
Binding Values	0.06 [-0.03, 0.15]	0.01 [-0.12, 0.14]	0.04 [-0.05, 0.14]	0.05 [-0.08, 0.18]
Individualizing Values	0.04 [-0.03, 0.12]	0.01 [-0.10, 0.11]	0.07 [-0.00, 0.16]	0.07 [-0.03, 0.18]

Note. Table shows the difference in effect of moral alignment vs misalignment or non-moral (neutral) framing on public sharing intentions for both false and true posts. Table shows no effect of moral alignment except across all conditions and for both true and false posts. Results indicate that private offline sharing is not driven by moral alignment.

9 Additional Models

Study 1

Table S18
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 1 hold when controlling for headline veracity.

Study 2

Table S19

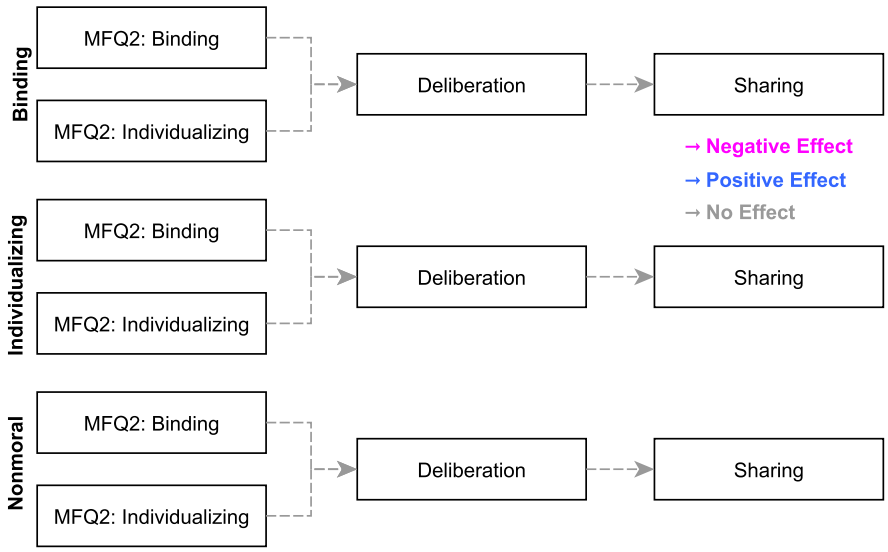
Effect sizes for Model 4 when controlling for headline veracity and familiarity

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 2 hold when controlling for headline veracity.

Figure S6

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.