The role of political devotion in sharing partisan misinformation

Clara Pretus
Universitat Autònoma de Barcelona https://orcid.org/0000-0003-2172-1184
Camila Servin-Barthet
Universitat Autònoma de Barcelona
Elizabeth Harris
William Brady
New York University
Oscar Vilarroya
Universitat Autònoma de Barcelona
Jay Van Bavel (jay.vanbavel@gmail.com)
New York University

Article

Keywords:

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1665189/v1

The role of political devotion in sharing partisan misinformation

Political Devotion & Misinformation

Clara Pretus[1,2,3], Camila Servin-Barthet[1,2], Elizabeth A. Harris[4], William J. Brady[5], Oscar Vilarroya[1,2], Jay J. Van Bavel[4*]

[1]Department of Psychiatry and Forensic Medicine, Universitat Autònoma de Barcelona, Spain.

[2]Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, Spain

[3]Center of Conflict Studies and Field Research, ARTIS International, St Michaels, MD.

[4]Department of Psychology and Center for Neural Science, New York University, New York, NY.

[5]Department of Psychology, Yale University, New Haven, CT.

***Corresponding author:**

Jay Van Bavel

Department of Psychology and Center for Neural Science, New York University

jay.vanbavel@nyu.edu

Abstract

Online misinformation poses a significant threat to global challenges, such as the COVID-19 pandemic and climate change. Misinformation is disproportionality shared by people with extreme political attitudes, especially among the far right. To understand the psychological and neurocognitive processes that underlie misinformation sharing among extremists, we conducted a pre-registered, cross-cultural experiment with conservatives and far-right supporters in the US and Spain (N = 1,609) and a neuroimaging study with far-right supporters in Spain (N = 36). Individuals who felt their personal identity was fused with their political group were more likely to share misinformation, especially when the misinformation was related to issues that involve sacred moral values (e.g., immigration and nationalism). Analytical thinking was unrelated to misinformation sharing when the misinformation involved sacred values (vs. non-sacred values) and fact-checks had little or no effect in this sample, especially among hyper-partisans. Far-right supporters also showed increased activity in brain regions associated with theory of mind

in response to posts with sacred values, highlighting the social dimension of misinformation sharing. These results suggest that political devotion plays a key role in misinformation sharing and that identity-based interventions may help curb misinformation for specific groups.

Introduction

As social media increasingly becomes our main source of news[1], malicious agents are using this opportunity to spread misinformation to larger audiences faster than ever before[2,3]. Attempts to quantify this phenomenon suggest that 70% of Twitter users are exposed to fake news[4], and fake news is 70% more likely to be shared than true news, consistently spreading more quickly and more broadly[5]. Critically, online misinformation can impact real-world outcomes such as fueling ideological polarization[6–8] and reducing vaccination intentions[9]. Thus, it is critical to understand the psychological processes associated with sharing online misinformation as well as potential interventions that may help counteract its spread.

One of the major frameworks currently employed to understand why individuals share misinformation is that they lack the skill or motivation to engage in analytical reasoning[10]. Several studies have found that a lower propensity to engage in analytical thinking is associated with poor truth discernment in news headlines and that accuracy nudges improve sensitivity to falsehood[11]. Other studies find no relationship between cognitive reflection and misinformation sharing[12] and the impact of accuracy nudges may be quite weak[13]. While correctly identifying a headline as false may stop analytical thinkers from sharing it, it does not explain what makes them *want* to share it in the first place. A recent meta-analysis of this body of work found that these accuracy nudges had little or no effect on misinformation sharing among far right-wing participants[14], suggesting that political ideology or identity may play a key role in the spread of misinformation (see also Guess et al.[15]).

In this respect, partisanship or identification with a political party has been identified as one of the main factors driving belief and willingness to share misinformation [12,16,17]. According to the identity-based model of political belief[18], individuals are more inclined to believe (mis)information that affirms their partisan identities; that is, beliefs that serve

belonging, epistemic, status and moral goals associated with their social identity. Thus, for accuracy nudges to work, the benefits of being accurate need to outweigh the partisan motives, or accuracy needs to be linked to a particular social identity (e.g., as with scientists). The model also makes key predictions about the neural processes underlying partisan motivations. Particularly, these value computations may be mediated by the orbitofrontal cortex[18], a richly interconnected brain region thought to integrate overall value during decision-making and generate evaluations[19,20].

The incentives of identity-motivated cognition may be especially relevant for hyper-partisans or individuals with extreme identities. Research on extremism often refers to extreme partisans as "devoted" actors[21], individuals who are characterized by two identity motives: identity fusion and willingness to sacrifice for sacred values. Identity fusion is a visceral feeling of oneness with a group that results from the merging between their social and personal identities[22]. Unlike group identification[23], which involves a process of depersonalization and adjustment to group norms[24], identity fusion allows preserving a strong sense of self. It is precisely this heightened sense of self that enables devoted actors to initiate extraordinary actions such as risking their lives on behalf of the group[22]. Moreover, devoted actors are more likely to engage in these behaviors when the values at stake are perceived to be sacred[21]. Sacred values are strongly held beliefs resistant to economic trade-offs[25,26] that often lie at the heart of intractable conflict[27], e.g., the "holy land" of Israel. The non-negotiable status of sacred values has been conceptualized as inherently linked to their identity-defining nature (e.g., "This is who we are and what we stand for", see Sachdeva et al.[28]). If devoted actors are willing to risk their lives to fight for sacred values, are they also willing to share misinformation online to promote their cause, even at the cost of factual truths?

We propose that spreading partisan misinformation can be understood as a consequence of political devotion. In line with this, we suggest that the two components of political devotion– identity fusion and sacred values–are critical identity motives that drive engagement with partisan misinformation. In terms of the identity-based model of political belief[18], identity fusion can be understood as a form of extreme partisanship, where sacred values are highly moralized beliefs that serve partisan goals. If identity motives play any role in misinformation sharing, then two consequences should follow. First, partisan goals should

be weighted more heavily than accuracy goals among extremists (e.g., fused individuals) and whenever highly relevant partisan goals are invoked (e.g., misinformation appeals to sacred values), which should lead to increased misinformation sharing. Second, when it comes to behavioral interventions, attempting to enhance accuracy concerns, for instance, by using fact-checks or accuracy nudges against misinformation, should be less effective when identity motives are most salient (e.g., in fused individuals and for misinformation on sacred values). Ultimately, sharing of partisan misinformation should be higher whenever and for whomever identity motives overshadow accuracy concerns.

Current research

In the present work we explore how the two fundamental identity motives of political devotion, sacred values and identity fusion, impact people's likelihood of sharing misinformation across two countries (the US and Spain). In this context, we evaluate whether these identity motives are associated with higher likelihood of sharing misinformation and reduced effectiveness of interventions aimed to increase accuracy concerns and reduce misinformation sharing. Specifically, we assessed the efficacy of different popular interventions, including *fact-checks* used by social media companies ("This claim about…is disputed."), *accuracy nudges* supported by widely cited papers ("To the best of your knowledge, is the above statement accurate?", based on Pennycook et al.[11]), and *media literacy nudges* ("What techniques are used in this Tweet to attract your attention?").

Because conservatives and far-right supporters disproportionally share far more misinformation[4,15,29], we conducted two pre-registered experiments with these populations in the US and Spain (N = 1,609). Importantly, because extreme partisanship is also possible in the political left, our research does not preclude that identity motives similarly influence misinformation processing in the far left. Moreover, to better understand the neurocognitive processes underlying misinformation sharing, we conducted a neuroimaging study with far-right supporters in Spain (N = 36). The neuroimaging study aimed to identify which brain networks are involved in processing misinformation related to sacred (*vs.* non-sacred) values, as well as to evaluate the brain response to the tested intervention. For instance, interventions aimed to reduce misinformation by increasing analytical thinking should be associated with increased activity in brain regions that support cognitive control such the dorsolateral

prefrontal cortex[30], or prediction error such as the anterior cingulate, which have been associated with subsequent behavioral adjustment[31]. Moreover, if an intervention is ineffective, the recorded brain response could provide clues as to why it is ineffective (e.g., the neurofunctional analysis suggests different networks are active). Thus, understanding the psychological motives and neuro-cognitive processes among extreme partisans will be critical to understand one of the key sources of online misinformation.

To understand the spread of misinformation, we focused on false statements in this research. We did not include true statements because we were primarily interested in testing the impact of using sacred moral values as a vehicle to spread misinformation, rather than trying to identify how people differentiate false from true information. Ultimately, misinformation in its most compelling form is only distinguishable from true information in that it does not correspond to reality. This choice has two implications: first, participants in this series of studies were exposed to an unrealistic volume of misinformation, considering that misinformation in the real world coexists with true information[15], and second, our findings do not preclude that appealing to sacred values is associated with higher likelihood of sharing any type of information, regardless of whether it is true or false.

In line with the idea that identity motives should outweigh accuracy concerns whenever highly relevant partisan goals are invoked, we predicted that misinformation relevant to sacred values (*vs.* non-sacred values) would be associated with greater likelihood of sharing, and that fused (*vs.* non-fused) individuals would be more likely to share partisan misinformation. In addition, previous literature suggests that behavioral interventions aimed to enhance accuracy concerns such as fact-checking and accuracy nudges should decrease the likelihood of sharing misinformation. However, because sacred values are highly relevant identity motives, they are particularly resistant to trade-offs and social influence[32]. Thus, we expected it to be harder for interventions focused on enhancing accuracy to override identity motives for misinformation relevant to sacred (vs. non-sacred values), leading to a reduced effect of fact-checking in this condition. In terms of brain activity, we expected the orbitofrontal cortex to be involved in sharing messages that include highly relevant identity motives or sacred values in line with the higher subjective value involved in these decisions. In terms of the brain response to the intervention, we expected to observe increased activity in brain regions associated with cognitive control or prediction error as explained above.

Results

The appeal of sacred values

In two survey studies with Spanish (N = 812) and US conservatives (N = 797) and a neuroimaging study with Spanish far-right supporters (N = 36) participants were asked to rate the likelihood of sharing a series of social media posts composed by ingroup political leaders (see demographic data in Table 1). The social media posts were specifically designed to convey 8 false statements about sacred political issues (immigration, nationalism, and women and family values) and 8 false statements about non-sacred political issues (roads and infrastructure, foreign affairs, waste management, and materials).

As predicted, the presence of sacred values was associated with higher likelihood of sharing misinformation across the three samples (Spanish sample: $M_{diff} = 0.54$, 95% CI [0.48, 0.60], $t(811) = 18.80$, $p < .001$, Cohen's d = 0.41, 95% CI [0.31, 0.51]; US sample: $M_{diff} = 0.56$, 95% CI [0.50, 0.63], $t(796) = 16.03$, $p < .001$, Cohen's d = 0.38, 95% CI [0.28, 0.48]; fMRI sample: $M_{diff} = 1.14$, 95% CI [0.82, 1.47], $t(34) = 7.09$, $p < .001$, Cohen's d = 1.06, 95% CI [0.58, 1.60], see Table 2a and Fig. 1a). Moreover, the impact of sacred values increased as a function of the number of sacred values held by participants (out of the three proposed ones, see Fig. 1b) in both Spanish conservatives (B = 0.18, 95% CI [0.13, 0.24], $t(810) = 6.60$, $p < .001$) and US conservatives (B = 0.19, 95% CI [0.13, 0.25], $t(795) = 6.41$, $p < .001$), though the effect of number of sacred values was in the same direction, but not statistically significant in the fMRI sample (B = 0.14, 95% CI [-0.13, 0.41], $t(33) = 0.97$, $p = .34$). Thus, participants across the three studies were consistently more likely to share misinformation about sacred issues (e.g., immigration) compared to non-sacred issues (e.g., infrastructure), especially if they personally held these values as sacred, and even if both types of issues were formulated as a critique to the government or the status quo.

To test for potential confounds, we pilot tested the social media posts employed in the US study in an independent sample of US conservatives (N = 80). We asked this sample to rate sacred and non-sacred values on perceived accuracy, familiarity, salience, and attitude strength (certainty, extremity, and importance; $\alpha = 0.87$). Social media posts relevant to sacred (vs. non-sacred) values received similar ratings in perceived accuracy (sacred values: M = 3.28, SD = 0.69; non-sacred values: M = 3.14, SD = 0.66; $M_{diff} = 0.14$, 95% CI [-0.01, 0.29], $t(72) = 1.80$, $p = .076$, Cohen's d = 0.21, 95% CI [-0.12, 0.53], see Fig. 1g), but were

rated higher in attitude strength (sacred values: M = 3.59, SD = 0.77; non-sacred values: M = 2.45, SD = 0.80; M_{diff} = 1.14, 95% CI [0.95, 1.34], t(72) = 11.72, p < .001, Cohen's d = 1.46, 95% CI [1.09, 1.82]), familiarity (sacred values: M = 3.25, SD = 0.88; non-sacred values: M = 2.17, SD = 0.96; M_{diff} = 1.09, 95% CI [0.89, 1.28], t(72) = 11.09, p < .001, Cohen's d = 1.18, 95% CI [0.83, 1.54]), and salience (sacred values: M = 2.60, SD = 0.97; non-sacred values: M = 1.86, SD = 0.80; M_{diff} = 0.74, 95% CI [0.57, 0.91], t(72) = 8.60, p < .001, Cohen's d = 0.83, 95% CI [0.49, 1.17]).

To assess and control the effect of each variable on participants' likelihood of sharing social media posts relevant to sacred (vs. non-sacred) values in our main US study, we reran the analyses accounting for the effect of each variable. Based on the pilot, we computed average scores in accuracy, attitude strength, familiarity, and salience for social media posts in the sacred versus non-sacred value condition for each participant in the US study. Because different participants had been exposed to a different set of social media posts (8 out of 32 possible items), confound scores also varied across participants. When assessed separately, all variables influenced likelihood of sharing social media posts (accuracy: B = 1.13, 95% CI [0.94, 1.32], t(950) = 11.67, p < .001; attitude strength: B = 0.78, 95% CI [0.67, 0.89], t(842) = 13.88, p < .001; familiarity: B = 0.79, 95% CI [0.68, 0.90], t(857) = 14.14, p < .001; and salience: B = 0.84, 95% CI [0.72, 0.95], t(867) = 13.83, p < .001). However, when added into the same model together with value sacredness, *value sacredness remained the only significant predictor of likelihood of sharing social media posts* (value sacredness: B = 0.62, 95% CI [0.44, 0.80], t(1091) = 6.78, p < .001; accuracy: B = 0.26, 95% CI [-0.15, 0.66], t(967) = 1.24, p = .216; attitude strength: B = 0.02, 95% CI [-0.45, 0.48], t(1004) = 0.07, p = .943; familiarity: B = 0.06, 95% CI [-0.36, 0.48], t(938) = 0.29, p = .772; and salience: B = -0.32, 95% CI [-0.77, 0.12], t(1140) = -1.42, p = .157). Since both value sacredness and the variable scores for each social media post were obtained from sample averages, differences in individual-level versus sample-level measurements did not appear to influence these results. Thus, sacred (vs. non-sacred) values were associated with increased likelihood of sharing social media posts above and beyond differences in attitude strength, familiarity, and salience.

The neuroimaging study revealed very strong brain activity in response to social media posts containing sacred values compared to non-sacred values among far-right

supporters. Due to generalized activation across the brain at a standard threshold ($p < .001$ FWE cluster-level) in this contrast, we had to use a more stringent statistical threshold ($p < .05$ FWE peak-level) to detect several prominent clusters of activity in the left middle temporal gyrus, left dorsomedial prefrontal cortex, left precuneus, left middle and inferior frontal gyrus, and the bilateral occipital cortex (thresholded at $T = 5.57$, $p < .05$ FWE, see Fig. 2A and Table 3a). Based on the degree of overlap between the observed brain activity pattern and the seven functional networks described in Yeo et al.[33], 56% of the identified activity to sacred values overlapped with the default mode network, 28% with the visual network, and 12% with the frontoparietal network (the remaining 4% was distributed across the somatomotor network, the dorsal and ventral attention networks, and limbic areas). A location analysis with *Neurosynth*[34] indicated that the peak activations in each cluster were broadly associated with terms related to language and social inferences (see Supplementary results) and the obtained pattern of activation considerably overlapped with a *Neurosynth* mask associated with *theory of mind* (see Fig. 2d).

Effects of political affiliation and identity fusion

Identity fusion was assessed in relation to the political party participants were affiliated with ("Vox" for far-right voters, "Partido Popular" for center-right voters, and "Republican party" for US conservatives). In addition, republicans in the US were also asked to rate their identity fusion with Trump (following Kunst et al.[35]). Unlike fusion with the Republican party, fusion with Trump has been found to predict willingness to engage in political violence[35]. Identity fusion was more frequent among the far-right in Spain (15%, 60 out of 408) and in relationship to Trump in the US (15%, 122 out of 797) compared to the Spanish center-right (6%, 35 out of 404) or the Republican party in the US (6%, 51 out of 797, were fused with the Republican party but not Trump and 14% were fused with both). Analyses of political affiliation and identity fusion effects were pre-registered as exploratory. Similarly, to Spanish far-right voters, conservatives fused with Trump had more sacred values and were more likely to share misinformation that included sacred values, compared to center-right voters and Republicans not fused with Trump, respectively (see Table 2b, 2f and Fig. 1C and 1D). Identity fusion was associated with a greater likelihood of sharing misinformation across groups (see Table 2b-2f).

Participants in the neuroimaging study were at the far-right end of the political spectrum (8.27 out of 10 points in a liberal to conservative scale, see demographics in Table 1), similar to US survey respondents fused with Trump (8.50 out of 10). However, only 5 of them were completely fused with the far-right party, a similar proportion than that found among far-right voters in the Spanish survey (14%). The reduced number of fused participants in the fMRI study precluded any comparisons between fused and non-fused individuals.

Limited efficacy of fact checking

In our first experiment with Spanish conservatives, we randomly assigned people into three conditions, each of which was fact-checked with a different statement: (1) a regular Twitter fact-check ("*This claim about…is disputed.*"), (2) an accuracy nudge ("*To the best of your knowledge, is the above statement accurate?*", based on Pennycook et al.[11]), and (3) a media literacy nudge ("*What techniques are used in this Tweet to attract your attention?*", based on content published by the Center of Media Literacy, *www.medialit.org*). All participants first completed a baseline block which did not include any fact-checks, followed by an experimental block that contained social media posts with and without fact-checks. Contrary to prior work, fact-checks did not have any overall effect within the experimental block across the three groups (M_{diff} = -0.05, 95% CI [-0.11, 0.02], t(1622) = -1.73, p = .20, Cohen's d = -0.04, 95% CI [-0.13, 0.06]) and the type of fact-check did not influence likelihood of sharing misinformation during fact-checked trials compared to control trials (Literacy nudge: p = .88, Cohen's d = -0.02, 95% CI [-0.19, 0.15]; Accuracy nudge: p = .89, Cohen's d = -0.02, 95% CI [-0.18, 0.15]; Twitter fact-check: p = .10, Cohen's d = -0.07, 95% CI [-0.24, 0.10], see Fig. 1f). Popular interventions aimed to reduce misinformation sharing by targeting accuracy and media literacy were thus notably ineffective among conservatives and far-right voters.

In light of these null effects, we only used the Twitter fact-check in the following US survey and fMRI study. Because Spanish respondents had increased their likelihood of sharing during fact-checked trials compared to the baseline block (M_{diff} = 0.16, 95% CI [0.10, 0.22], t(1622) = 5.87, p < .001, see Fig. 1e) we decided to add a control group with no fact-checks in the US survey. This allowed us to test whether this increase was a temporal effect

or a backfire effect of fact-checking (see Nyhan and Reifler[36]). Comparing the two groups confirmed the temporal effect: while the control group was more likely to share social media posts in the second experimental block compared to the baseline block (M_{diff} = 0.14, 95% CI [0.06, 0.23], t(1173) = 3.33, p < .001, Cohen's d = 0.10, 95% CI [-0.04, 0.24]), the experimental group did not increase likelihood of sharing in the experimental block compared to baseline (M_{diff} = 0.01, 95% CI [-0.06, 0.09], t(1173) = 0.37, p = .71, Cohen's d = 0.01, , 95% CI [-0.11, 0.13], see Fig. 1f). Thus, fact-checks had a small effect keeping participants from increasing sharing in the experimental block.

Contrary to predictions, type of value (sacred vs. non-sacred) did not moderate the effect of fact-checks in neither experiment (Spanish sample: sacred values p = .20, Cohen's d = -0.04, 95% CI [-0.14, 0.05] and non-sacred values p = 0.76, Cohen's d = -0.02, 95% CI [-0.11, 0.08]; US sample: sacred values p = .03, Cohen's d = -0.08, 95% CI [-0.22, 0.06] and non-sacred values p = 0.03, Cohen's d = -0.09, 95% CI [-0.23, 0.05]).

An exploratory analysis revealed that identity fusion with Trump in the US sample moderated the effect of fact-checks in the fact-checked group. Conservatives who were not fused with Trump decreased sharing during fact-checked compared to control trials (M_{diff} = -0.19, 95% CI [-0.29, 0.09], t(376) = -3.74, p < .001, Cohen's d = -0.13, 95% CI [-0.29, 0.02]), but conservatives fused with Trump did not (M_{diff} = -0.02, 95% CI [-0.24, 0.19], t(376) = -0.21, p = .84, Cohen's d = -0.01, 95% CI [-0.35, 0.32], see Fig. 1f). Hence, while Twitter fact-checks were somewhat effective in reducing misinformation sharing among non-fused republicans, they were ineffective among highly identified Trump supporters.

To explore possible psychological processes underlying the reduced effect of fact checks in participants fused with Trump, we ran an exploratory analysis assessing how these participants responded to fact checks in social media posts with different levels of perceived accuracy. To maximize differences between social media posts associated with high and low perceived accuracy, we selected only the top 10 and the bottom 10 social media posts (out of 32 displayed posts) which were associated, respectively, with the highest (M = 3.96, SD = 0.19) and lowest (M = 2.71, SD = 0.29) perceived accuracy ratings obtained in the US pilot (N = 80). We then ran a model comparing the effects of fact-checking on the likelihood of sharing high (vs. low) perceived accuracy social media posts in participants fused with

Trump (vs. non-fused). We did not find any evidence of a three-way interaction (p = .546) but given the relatively low statistical power we decided to analyze simple effects. Specifically, we found that fact-checking was effective only for low perceived accuracy posts in participants who were not fused with Trump (M_{diff} = -0.30, 95% CI [-0.47, -0.12], t(905) = -3.35, p < .001, Cohen's d = -0.19, 95% CI [-0.36, -0.02], see Fig. 1i). Conversely, fact-checks were ineffective for low perceived accuracy posts in participants fused with Trump (M_{diff} = -0.20, 95% CI [-0.58, 0.18], t(909) = -1.02, p = .307, Cohen's d = -0.09, 95% CI [-0.46, 0.29]) and for high perceived accuracy posts across groups (Fused with Trump: M_{diff} = 0.17, 95% CI [-0.23, 0.57], t(918) = 0.83, p = .405, Cohen's d = 0.09, 95% CI [-0.29, 0.48]; non-fused with Trump: M_{diff} = -0.12, 95% CI [-0.30, 0.06], t(909) = -1.29, p = .196, Cohen's d = -0.05, 95% CI [-0.23, 0.12], see Fig. 1h). Thus, our exploratory analysis tentatively suggests that fact-checks are most effective for low perceived accuracy posts, though not for people fused with Trump.

In the fMRI study, participants saw the Twitter fact-check in half of the trials. Though the effect size of the Twitter fact-check was similar to that found in the US and Spanish surveys, the effect was not statistically significant with a sample of 35 participants (M_{diff} = -0.06, 95% CI [-0.16, 0.04], t(34) = -1.28. p = .21, Cohen's d = -0.07, 95% CI [-0.54, 0.41]). At a neural level, we expected interventions aimed to reduce misinformation sharing to increase activity in brain areas associated with cognitive control or with prediction error, which have been associated with subsequent behavioral adjustment. However, the interaction between sacred values and fact-checking revealed a network comprising the right dorsomedial prefrontal cortex, the right precuneus, the left hippocampus and the bilateral cerebellum cortex; thresholded at T = 3.36, k = 298, $p < .001$ FWEc. At baseline, these areas were more active during the sacred value condition, but they became similarly active for sacred and non-sacred values when they were fact-checked (see Fig. 2b and Table 3b). 91% of this brain activity pattern overlapped with the default mode network, while the rest of the activity mostly overlapped with the visual network (8%). A meta-analysis of locations using *Neurosynth*[34] revealed associations between the obtained cluster peaks and terms related to theory of mind and autobiographical recall (see Supplementary results). Thus, rather than eliciting activation in brain regions associated with cognitive control and behavioral

adjustment, fact-checks resulted in increased activity in brain areas associated with theory of mind.

Neural activity in similar areas was higher in participants who started getting fact-checked in the second half of the fMRI session compared to those who were fact-checked from the beginning (see Supplementary results, Fig. 2c and Table 3c), suggesting a decreased response to fact-checking in participants who were habituated to fact-checks.

Cognitive reflection and other variables of interest

Analytical thinking style, as measured by correct responses in the cognitive reflection test, has been previously found to predict reduced sharing of fake news[10]. However, we found that, while analytical thinking style was associated with reduced sharing of misinformation related to non-sacred values (M_{diff} = -0.29, 95% CI [-0.57, -0.003], t(982) = -1.98, p = .047, Cohen's d = -0.16, 95% CI [-0.43, 0.09]), it did not affect sharing of misinformation related to sacred values (M_{diff} = -0.04, 95% CI [-0.32, 0.24], t(982) = -0.28, p = .78, Cohen's d = 0.08, 95% CI [-0.18, 0.34], see Fig. 1e). As such, the cognitive reflection test was not administered in the US survey. Strategies based on analytical thinking may not be very effective at reducing sharing of fake news that directly targets sacred issues such as immigration.

With regards to other variables of interest, both media literacy and humility, but not scientific curiosity, were associated with increased sharing of posts with sacred values compared to non-sacred values (see Supplementary results). The presence of moral-emotional language increased sharing in the Spanish sample, while in the US sample it only increased sharing among Republicans fused with Trump (see Supplementary results).

Discussion

We investigated the role of sacred values and identity fusion in likelihood of sharing political misinformation in conservatives and far-right supporters in Spain and the US. Across three experiments, we found that appealing to sacred values in political messages increases likelihood of sharing misinformation on social media. The effect of sacred values was particularly strong in people who have a fused sense of identity with the far-right in Spain and with Donald Trump in the US. Moreover, we found that a variety of fact-checks

employed by social media companies and supported by widely cited papers (e.g., Pennycook et al.[11]) were largely ineffective – especially in individuals fused with Trump. Similarly, whereas analytical thinking was associated with lower sharing of misinformation about non-sacred values (e.g., infrastructure), it was irrelevant for misinformation around sacred issues (e.g., immigration). In terms of brain activity, we observed a very strong increase in regions that highly overlapped with *theory of mind* networks in response to messages that included sacred values (vs. non-sacred values).

Our findings are consistent with a significant role of partisan identity in misinformation sharing[37], which stems from the idea that individuals are motivated to identify and conform with groups that provide them protection, status, and resources[38]. As they identify and bond with the group, individuals engage in identity motivated cognition, processing information in a way that allows them to affirm their identity[18,39,40]. Thus, partisans appear motivated to believe and share identity-relevant information, with little regard for accuracy[18]. Accordingly, we found that, while analytical thinking somewhat reduced sharing of misinformation about identity-irrelevant topics (i.e., non-sacred values such as infrastructure), it made no difference for identity-relevant topics that tackled highly cherished conservative values, like immigration, nationalism, and traditional family values. These results suggest that individuals have a strong drive to share identity-relevant information regardless of how analytical they may be. Several other studies find that partisan motives override accuracy concerns when it comes to sharing news online. For instance, in a study combining survey-based questionnaires and Twitter profile analysis party identification was the strongest predictor of fake news sharing, especially among Republicans, while analytical thinking and media literacy were not significant[12]. Party identification was also the strongest predictor of believing fake news[16]. Among partisans on the right wing, traditional attempts to fact-check them or increase analytic thinking appear to be less effective than among other populations.

Overall, sharing of misinformation was consistently higher for sacred values across samples, even after controlling for attitude strength, familiarity, and salience. This result suggests that malevolent agents trying to spread misinformation benefit from framing false information in terms of sacred values. Consequently, social media users should be particularly careful in discerning the truthfulness of new information that is presented in this

format. The identity-defining nature of sacred values helps explain their remarkable appeal when it comes to misinformation sharing. It is important to note that not every value that is *identity-relevant* is also *identity-defining* (e.g., sacred), and new sacred values can emerge depending on how leaders frame relevant issues as they become salient. For instance, at the time the US was contemplating sanctions against the Iranian nuclear program in 2009, the Iranian government framed the nuclear program as an "inalienable right" and many Iranians started perceiving it as a sacred issue[41]. Surely, sacred values may be intertwined with elite interests, group interests and ideological concerns. However, disentangling how these constructs are interrelated falls outside of the scope of the present work. Instead, our main goal was to evaluate how non-negotiable identity-defining values are optimally suited to function as vehicles to deliver misinformation.

Critically, the effect of sacred values on misinformation sharing was even more prominent for individuals who were fused with extreme political options such as the far-right in Spain and Trump in the US. The fact that highly committed individuals were more willing to share misinformation supports the notion that identity motives more readily override accuracy concerns in extreme partisans, which is unsurprising when considering previous literature on extremism. For instance, previous work found that Moroccans who deemed sharia law a sacred value and were fused with a kin-like group supported militant jihad and were willing to engage in costly sacrifices for the implementation of sharia[42]. In the US context, fusion with Donald Trump predicted willingness to violently prosecute Muslims, immigrants and personally protect the US border[35]. If partisan incentives can be high enough for extremists to be willing to risk their lives, it is unsurprising that they can also outweigh accuracy concerns.

At a brain level, we found a strong neural response to messages involving conservative sacred values (vs. non-sacred values) in far-right supporters. This activity highly overlapped with brain regions typically associated with social cognition and theory of mind, such as the dorsomedial prefrontal cortex, the middle temporal gyrus and the precuneus[43,44]. These neural findings provide additional support to the notion that identity motives, which are linked to social-cognition processes, are critical in processing the type of misinformation that is more likely to be shared (i.e., misinformation that tackles sacred values). A similar neural activation pattern increased for sacred and non-sacred values that

were being fact-checked, especially for participants who were unhabituated to fact-checks. These results suggest that socio-cognitive processes are not only at the root of misinformation sharing, they are also engaged in processing fact-checks; that is, when participants come across fact-checks, they seem to mobilize theory of mind networks (for instance, for assessing intentions) rather than brain regions that support analytical thinking such the frontoparietal network. While the present design does not allow us to infer what exact socio-cognitive processes our participants were engaged in, there are several plausible options that would be interesting to test in future studies. One option is that they are evaluating the appropriateness of sharing the post by gauging on how others would judge it (especially the ingroup). This process would be particularly relevant for social signaling. Alternatively, because identity and self-referential processes neurally overlap with social cognition, especially in the medial prefrontal cortex[45], participants could be strengthening their partisan identities by incorporating new (false) information about their ingroup's worldview.

The immediate consequence of misinformation sharing being a fundamentally social process is that popular interventions such as fact-checking and accuracy nudges are less likely to be effective among highly fused individuals. Prior work has shown that respondents are more sensitive to falsehood in news headlines after being asked to rate their accuracy[11]. However, we find close to no effect of fact-checking in our conservative samples, especially for fact-checks based on accuracy and media literacy. Fact-checks were particularly ineffective in participants fused with Trump, suggesting that, relative to accuracy incentives, partisan incentives become even more relevant for hyper-partisans.

One possible explanation for these null effects is that the stimuli we designed were too plausible. Previous studies have found that the disposition for analytical thinking is unlikely to help reduce belief in misinformation that is plausible[10]. Thus, it could be that fact-checks and nudges are helpful specifically when the presented misinformation is clearly implausible. We tested this idea in an exploratory analysis comparing the effect of fact-checks on high and low perceived accuracy items in the US sample. Because asking participants to make accuracy judgements could influence subsequent likelihood of sharing[11], we obtained accuracy judgements for each social media post in a separate pilot study using a similar sample. Our results are consistent with the notion that fact-checks are most useful in reducing sharing of misinformation that is perceived to be less accurate on average.

However, this effect was absent in participants strongly fused with Trump, who still did not decrease sharing of fact-checked posts containing claims deemed to be implausible by other Republicans. Thus, there appears to be an important role for social identity in the efficacy of these interventions (see also Rathje et al.[14]). Future research should test this hypothesis with larger samples of individuals fused with Trump to confirm these preliminary results.

The limited effectiveness of fact-checks on conservatives and, particularly, Republicans fused with Trump is in line with other studies showing that conservatives and those at the far-right of the political spectrum are less sensitive to false claims on social media[29], are more exposed to them[4] and six to seven times more likely to share them than moderates and liberals[4,15]. However, because conservatives and liberals perform equally well on the cognitive reflection test[39], lack of analytical reasoning cannot explain the differences in misinformation sharing and susceptibility to fact-checks across the political spectrum. Our data suggest that extreme political movements have a greater capacity to mobilize partisans and fuse them with the movement. Identity fusion rates for the far-right in Spain and Trump in the US are notably higher (15% of the sample) than for the center-right in Spain (9%) and the Republican party (6% exclusively fused with the party but not Trump). Because identity fusion predicts misinformation sharing and susceptibility to fact-checks, movements' capabilities to elicit identity fusion among supporters could contribute to political asymmetries in these two dimensions.

In sum, our work emphasizes the role of political devotion in fueling misinformation sharing and suggests that socio-cognitive processes may be critical for the spread of misinformation among partisans. While accuracy nudges may work for non-partisans (albeit with very small effect sizes), these populations are probably not at the root of the misinformation crisis, because they are already less prone to sharing misinformation. The speed and ease with which fake news spread as compared to true news[5] may be a result of how this misinformation is designed. Fake news creators likely benefit from tackling sacred issues, thus mobilizing large crowds of who hold a sincere belief that they are fighting for a legitimate cause. For instance, InfoWars, one of the most popular websites known to spread fake-news, mobilizes their followers by calling them into *battling globalism*, an idea that may be powerful for many. By appealing to a cause greater than the self, identity-relevant misinformation becomes very tempting, even for highly analytical partisans. Spreading

misinformation might be thus a warfare tool employed by interest groups to amplify their messages and gain power through devoted partisans.

The present work is subject to several limitations. Because of previously reported political asymmetries in information sharing[4,15,29], we decided to focus on conservatives. However, our data indicate that extreme political options elicit higher levels of identity fusion among supporters, suggesting that far-left supporters could possibly show similar patterns of misinformation sharing relative to sacred values. Moreover, although the fMRI sample was politically extreme (their average political orientation was 8.27/10, similarly to participants fused with Trump), it did not include enough fused participants to be able to test the effect of identity fusion in the scanner. Similarly, we did not include a control group in the fMRI study, so that the presented results are not necessarily specific to far-right supporters. In fact, we believe we could find similar neural results in other groups exposed to misinformation relevant to their own sacred values. For instance, a recent nationally representative sample of Americans found that people on the extreme left score higher in dogmatism (albeit not as far as those on the far right) and they may therefore be resistant to information that challenges their beliefs (see Harris et al.[46]). Finally, our study was a controlled experiment with artificially designed social media posts. Future work will benefit from field studies looking at the relationship between identity fusion, sacred values, and fact-checks in real social media ecosystems.

Conclusion

Overall, our findings suggest that stopping the spread of misinformation may be a much greater challenge than previously thought. Our data suggests that believing and sharing fake news is, at least in part, a socio-cognitive process serving partisan identity goals[18]. Misinformation is specifically designed to tackle sacred beliefs that forge people's partisan identity and to which they hold on with great conviction. Critically, individuals' willingness to spread this type of content appears resistant to analytical thinking and nudges aimed to improve accuracy, especially among hyper-partisans, who are most likely to share it in the first place. Strategies aimed to reduce misinformation should thus address the devotional aspect of misinformation sharing.

Materials and Methods

This research was approved by the Ethics Committee on Human and Animal Experimentation at the Universitat Autònoma de Barcelona according to the Declaration of Helsinki guidelines (Ref. 5385 and 5388). The pre-registration of this study can be found following this link.

Experimental Design

We launched surveys in Spain (Study 1, N = 812) and the US (Study 2, N = 797), and conducted a neuroimaging study (N = 36) asking participants to rate the likelihood of sharing a series of social media posts composed by different political leaders and public figures who are part of, or openly support, the party they voted for in the last elections. Because the Spanish congress has multiple parties, we chose to focus on voters of two conservative political parties in Spain (center-right and far-right) in Study 1, voters of Republican Donald J. Trump in the US in Study 2, and voters of the far-right in Spain for the neuroimaging study.

In the three studies, half of the posts, designed to look like tweets, included conservative sacred values (immigration, nationalism, and women and family values) and the other half non-sacred values (roads and infrastructure, foreign affairs, and waste management and materials). Because previous studies suggest an effect of moral-emotional language in online content sharing (*43*), all items were formulated twice, once using moral-emotional language and once using neutral language (see further details in *Materials*). To confirm that these issues are sacred in Spain and the US, we assessed how many of our respondents deemed each value sacred in the three studies. To test the effect of online fact-checking on social media posts, we also introduced fact-checks in some of the social media posts. In Study 1 and 2, participants were first asked to rate four Tweets without fact-checks (baseline block) and then rate twelve (Study 1) or eight more posts (Study 2) in a mixed block with half of the Tweets marked with a fact-check (experimental block).

In Study 1 (N = 812) participants were split in three groups and each group was exposed to a different fact-check in the experimental block: a classic Twitter fact-check (N = 270, "*This claim about…is disputed.*"), an accuracy-based fact-check (N = 277, "*To the best of your knowledge, is the above statement accurate?*", based on Pennycook et al.[11]) and a media literacy-based fact-check (N = 265, "*What techniques are used in this Tweet to attract your attention?*", based on content published by the Center of Media Literacy, *www.medialit.org*). In Study 2 (N = 797) half of the sample was exposed to the classic Twitter fact-check and the other half did not see any fact-checks in any of the posts.

558 In the neuroimaging study (N = 36), an event-related fMRI paradigm was designed using
559 MATLAB 2020a (The MathWorks, Inc., Natick, Massachusetts, USA) with Psychtoolbox
560 extensions[47]. Participants completed two 6-minute runs without any fact-checks (baseline block)
561 followed by another two 6-minute runs with fact-checks in each post (experimental block). The order
562 of these two blocks was reversed in half of the sample (N = 18). Stimuli were presented randomly
563 with jittered inter-trial intervals that varied randomly from 0.5 s up to 3.70 s. Each run included 40 7-
564 second trials (20 trials for the sacred value condition and 20 trials for the non-sacred value condition)
565 and in each trial, participants were given 4.5 seconds to read a social media post, and 2.5 seconds to
566 respond how likely they would be to share it using a 6-point Likert scale adapted for fMRI using
567 left/right buttons (see Fig. S2). The orientation of the scale was reversed in half of the trials and the
568 cursor starting point was randomized. Half of the presented social media posts included sacred values
569 and the other half non-sacred values. The total duration of the paradigm was 24 minutes. Participants
570 had the chance to practice the task thoroughly on a laptop before going into the scanner. No posts
571 with fact-checks or sacred values were shown during the training phase.

572 ***Participants.*** Demographic data of each study sample is presented in Table 1. For Study 1,
573 408 Spanish far-right voters and 404 center-right voters were recruited by means of an online panel.
574 Inclusion criteria included being 18 years old or older and having the intention to vote for either a far-
575 right party ("*Vox*") or a center-right party ("*Partido popular*") in the next presidential election. Study
576 2 included a sample of 797 participants, recruited via an online panel, who had reported voting for
577 Republican D. J. Trump in the two previous US presidential elections (2016 and 2020).

578 A power analysis conducted with the R package 'simr'[48] revealed that, based on 1000
579 simulations, recruiting 800 participants would enable the detection of small effects (e.g. a slope of
580 0.1) with an alpha = 0.05 for the interaction term between (a) within-subjects value type (sacred vs.
581 non-sacred) and fact-checking (fact-checked vs. control) with a statistical power of 99.6% (95% CI
582 [98.98, 99.89]) in the Spanish and US survey with four observations per participant, (b) within-
583 subjects fact-checking (baseline, control and fact-checked trials) and fact-checking group (Twitter vs.
584 Accuracy vs. Literacy fact-check) with a statistical power of 79.1% (95% CI [76.45, 81.58]) in the
585 Spanish survey with three observations per participant and between (c) within-subjects block
586 (baseline vs. experimental block) and fact-checking group (fact-checked group vs. control group)
587 with a statistical power of 82.9% (95% CI [80.42, 85.18]) in the US survey with two observations per
588 participant.

589 For the neuroimaging study, a power analysis based on 1000 simulations and an alpha = 0.05
590 showed that with 40 participants and 80 observations per participant (40 trials per run in 2 runs), we

would be able to detect small effects (e.g., a slope of 0.1) for the interaction between within-subjects value type (sacred vs. non-sacred) and within-subjects fact-checking (control vs. fact-checked trials) with a statistical power of 99.80% (95% CI [99.28, 99.98]).

We recruited thirty-six far-right supporters for the neuroimaging study using social media. Initially, we contacted Twitter followers of far-right accounts based in Barcelona via direct message and recruited them to participate in the study. Ultimately, most of our participants were recruited through other participants that shared our study ad in far-right youth WhatsApp groups that are used to organize political actions locally and usually include a few hundred members. Participants were selected based on their response to the question "Which political party best represents your values and beliefs at this moment?". Only participants who responded with a far-right party ("*Vox*") and were 18 years old or older were recruited for the study. Candidates with neurological and psychiatric disorders, taking psychiatric medication, claustrophobic, or with metal parts in their body incompatible with an MR scanner were excluded.

Informed written consent was obtained from all participants at the start of each study. Participants were debriefed at the end of each study and were given the chance to discuss and ask questions about the study.

Materials. *Social media posts*. The Tweets were specifically designed to convey 8 false statements about sacred political issues (immigration, nationalism, and women and family values) and 8 false statements about non-sacred political issues (roads and infrastructure, foreign affairs, waste, and materials). All posts were designed by researchers to contain false information and most posts contained some sort of critique of the current liberal government both in Spain and the US (see complete list of items employed in the US study in Table S2). Some examples included:

> *3 out of 4 illegal immigrants who enter our country end up in criminal gangs, endangering our American way of life.* [sacred value; moral-emotional wording]

> *That the government has slashed the funding for American embassies in the Asia-Pacific region is an attack against the nation.* [non-sacred value; moral-emotional wording]

Social media posts with sacred (vs. non-sacred) values were combined with similar number of likes (M_{diff} = 2.50, 95% CI [-12.6, 17.6], t(14) = 0.35, p = .73), number of retweets (M_{diff} = -21.5, 95% CI [-120, 76.6], t(14) = -0.47, p = .64) and gender of the leaders tweeting the posts (the same 3 female and 4 male leaders tweeted the posts in each condition). The selection of leaders was different for each sample so that each voters' group had their own set of 7 ingroup leaders (center-right leaders, far-right leaders, and pro-Trump leaders). Nonetheless, all sets of leaders tweeted the exact same content (center-right and far-right leaders) or American adapted content (pro-Trump leaders).

Items in the sacred and non-sacred value conditions were also matched for character length in the Spanish version for Study 1 (M_{diff} = -1.69, 95% CI [-3.78, 0.40], t(30) = -1.65, p = .11), in the English version for Study 2 (M_{diff} = 2.44, 95% CI [-7.33, 12.2], t(30) = 0.51, p = .61) and in a shorter Spanish version adapted for the neuroimaging study (M_{diff} = 0.75, 95% CI [-1.31, 2.81], t(30) = 0.74, p = .46).

In addition, type of language was counterbalanced across both sacred and non-sacred value conditions; that is, all 16 items were formulated twice, once using moral-emotional language and once using neutral language. Two pilot studies (N = 45 each) were conducted to ensure that the two language formulations were perceived as different in terms of moral and emotional language in both the Spanish and the English versions. Results of these pilot studies can be found in supplementary materials (see Fig. S1). The final set of posts included 32 items.

To make sure that the sacred values we tackled were indeed sacred by our study samples, the value sacredness of each issue was measured in the three studies, as indicated in the *Measures* section. As expected, the three proposed sacred values were sacred by a large segment of each of the three samples as compared to the non-sacred values (see Table S1).

Finally, to control for the effect of potential confounds associated with sacred values in each of the social media posts, we ran a third pilot study on a sample of 80 US conservatives who were pre-screened using the same criteria as the US study participants (having voted for Trump in the two previous US presidential elections in 2016 and 2020). For each of the 32 social media posts employed in the US study, we obtained ratings of perceived accuracy ("To the best of your knowledge, is the claim in the above social media post accurate?" from 1 = "Very inaccurate" to 5 = "Very accurate"), certainty ("I feel very certain about my feelings toward this issue"), extremity ("I feel strongly about this issue"), importance ("This issue is personally important to me"), familiarity ("I am quite familiar with this issue"), and salience ("I spend a lot of time thinking about this issue") on a 5-point Likert scale from "Strongly disagree" to "Strongly agree".

Measures. *Likelihood of sharing.* Disposition to share each of the social media posts was assessed using a probability scale "If you were to see the above post on social media, how likely would you be to share it?"[11]. Responses could range from 1 "Extremely unlikely" to 6 "Extremely likely".

Value sacredness. Value sacredness was assessed by means of economic trade-off scenarios. We asked participants if they would be "*willing to give up* [value] *if that involved a great benefit to [American/Spanish] families such as a better economy, more jobs, better schools and hospitals, and in general, a better quality of life for all [American/Spanish] families including yours*" (following

657 Hamid et al.[49]. For instance, participants were asked whether they would allow the free entry of
658 immigrants in their country if that involved a great benefit for American families. Possible responses
659 included "Yes", "No" and "Maybe". A value was considered sacred if participants refused to give up
660 that value.

661 *Identity fusion.* Identity fusion was assessed using the pictorial measure[22], which includes
662 five pairs of circles with different degrees of overlap representing the relationship between the
663 participant (small circle) and the group (big circle). Respondents were asked to convey which pair of
664 circles best represents their relationship with the group. Identity fusion was assessed in relation to the
665 political party participants supported ("Vox" for far-right voters, "Partido Popular" for center-right
666 voters, and "Republican party" for Trump supporters). In addition, Trump supporters were also asked
667 to rate their identity fusion with Trump specifically (following Kunst et al.[35]).

668 *Scientific curiosity.* Scientific curiosity was assessed by means of three items extracted from
669 the stretching subscale of the Science Curiosity in Learning Environments scale[50], including "I try to
670 learn as much as I can in new situations", "I see a challenge as a way to grow and learn", and "I apply
671 new information to an existing problem to see if that helps". Participants responded using a five-point
672 Likert scale from 1 "Not at all like me" to 5 "Very much like me". The three items were averaged
673 into a Scientific curiosity score ($\alpha = .77$) with values ranging from 1.33 and 5 ($M = 4.07$, $SD = 0.67$).

674 *Intellectual humility.* Intellectual humility was evaluated with the 6-item open mindedness
675 subscale of Alfano et al.[51] intellectual humility scale. This subscale includes items such as "I think
676 that paying attention to people who disagree with me is a waste of time." and "I feel no shame learning
677 from someone who knows more than me.". The 5-point response scale ranged from 1 "Not at all like
678 me" to 5 "Very much like me". The six items were averaged into an Intellectual humility score ($\alpha =$
679 $.67$) with values ranging from 2.17 and 5 ($M = 1.12$, $SD = 0.60$).

680 *Media literacy.* Media literacy was measured by means of three items of the Media literacy
681 education scale by Simons and Meeus[52], including "I know that media represents information in a
682 selective way and know how to interpret media messages" and "I can evaluate media content taking
683 into account various criteria (e.g. accuracy of information, comparison of information, appreciation
684 of aesthetic aspects)", using a five-point Likert scale from 1 "Not at all like me" to 5 "Very much like
685 me". The three items were averaged into a Media literacy score ($\alpha = .70$) with values ranging from 1
686 and 5 ($M = 3.74$, $SD = 0.76$).

687 *Cognitive reflection test.* Analytical (vs. intuitive) thinking styles were assessed by means of
688 the cognitive reflection test[53] in the Spanish sample. This questionnaire includes three mathematical
689 problems such as "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How

much does the ball cost?". Correct answers were added into an Analytical thinking score with values ranging between 0 and 3 (M = 0.67, SD = 0.96).

Neuroimaging data. Magnetic resonance images were obtained in a Philips 3 T scanner. T1-weighted images were acquired using a FSPGR sequence (TR: 9.9 ms, TE: 4.6 ms, FA: 8, matrix size: 240 × 240, 180 slices, and slice thickness: 1.00 mm) and functional volumes were obtained using a EPI-T2* sequence (TR: 1750 ms, TE: 35 ms, FA: 70, matrix size: 76 × 76, 46 slices, and slice thickness: 3.1 mm).

Statistical Analysis

The data and code employed in the analyses are available following this link. Due to privacy concerns, the neuroimaging data can be obtained from the authors upon request with an approved IRB (we will send a password protected link to scientists with ethical approval).

Behavioral analyses. To evaluate the effect of sacred values, political affiliation, identity fusion and fact-checking in the Spanish and US surveys we conducted a series of mixed effects models with random intercepts for participants using REML (afex package in R[54]). To assess differences in the effect of sacred values on likelihood of sharing (a) between far right and center right voters in Spain and (b) between conservatives fused and non-fused with Trump in the US, we added an interaction term for group (a or b) and type of value (sacred vs. non-sacred value). Of note, the identity fusion pictorial measure (*22*) was analyzed as a dichotomous variable following previous literature, with the maximal score coded as "fused" and the rest as "non-fused".

To test the effect of fact-checks in trials with and without fact-checks in both surveys, we created two mixed effects models with interaction terms for (a) fact-check group (Twitter fact-check, accuracy fact-check and literacy fact-check) and presence of fact-check (baseline trials, fact-checked trials, and control trials) for the Spanish sample, and (b) fact-check group (Twitter fact-check vs. no fact-check/control group) and presence of fact-check (baseline block vs. experimental block) in the US sample. All employed tests were two-sided.

Neuroimaging analyses. Image processing was conducted using SPM12 (Wellcome Trust, University College London, UK) on MATLAB. Functional images were realigned and co-registered to the structural images, which were then segmented into white matter, gray matter, and cerebrospinal fluid. The forward deformation fields generated during the segmentation of the structural images were used to normalize the functional images. Finally, images were smoothed using an 8-mm FWHM kernel.

The first-level general linear model (GLM) included two regressors for stimulus presentation (for sacred vs. non-sacred value trials) with a duration of 4.5 seconds each, and two regressors for the response period (for sacred vs. non-sacred value trials) with a duration of 2.5 seconds each. The response regressors were parametrically modulated by number of keypresses in each trial. Six non-convolved head movement regressors were also included. In addition, first level GLMs included scrubbing regressors that accounted for outlier volumes as required by each participant. Outliers included volumes with a framewise displacement larger than 0.9mm or global BOLD signal changes above 0.5 SD. One participant had to be eliminated due to excessive head movement.

Group-level effects of sacred vs. non-sacred values during stimulus presentation time-locked to the start of the onset stimulus were evaluated by means of a t-test on individual level contrasts. Group differences related to the order of the runs (half of the sample started with fact-checks and the other half without fact-checks) were assessed using a two-sample t-test on individual level contrasts. The fact-check by value interaction contrast was computed by assigning counterbalanced weights to fact-checked (fc) vs. control (ctr) posts with sacred values vs. non-sacred values ([1 x ctr sacred values -1 x ctr non-sacred values -1 x fc sacred values 1 x fc non-sacred values]). *Post hoc* analyses included parameter estimate extractions of the cluster of activity in the dorsomedial prefrontal cortex in the fact-check by value interaction contrast and the group by value interaction contrast using Marsbar. Reported results were corrected for multiple comparisons using family-wise error (FWE) correction at cluster level, with a peak-level contrast of $p < .001$.

References

1. Gottfried, J. & Shearer, E. Americans' online news use is closing in on TV news use. *Pew Research Center* (2017).
2. Allcott, H. & Gentzkow, M. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* **31**, 211–36 (2017).
3. Nyilasy, G. Fake news: When the dark side of persuasion takes over. *https://doi.org/10.1080/02650487.2019.1586210* **38**, 336–342 (2019).
4. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
5. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science (New York, N.Y.)* **359**, 1146–1151 (2018).
6. Hang Au, C., Ho, K. K. W. & Chiu, D. K. W. The Role of Online Misinformation and Fake News in Ideological Polarization: Barriers, Catalysts, and Implications. *Information Systems Frontiers 2021* 1–24 (2021) doi:10.1007/S10796-021-10133-9.
7. Lee, F. L. F. Impact of social media on opinion polarization in varying times: *http://dx.doi.org/10.1177/2057047315617763* **1**, 56–71 (2016).
8. Spohr, D. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *http://dx.doi.org/10.1177/0266382117722446* **34**, 150–160 (2017).

9. Loomba, S., Figueiredo, A. de, Piatek, S. J., Graaf, K. de & Larson, H. J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour 2021 5:3* **5**, 337–348 (2021).
10. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
11. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D. G. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention: *https://doi.org/10.1177/0956797620939054* **31**, 770–780 (2020).
12. Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A. & Petersen, M. B. Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review* 1–17 (2021) doi:10.1017/S0003055421000290.
13. Roozenbeek, J., Freeman, A. L. J. & Linden, S. van der. How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020): *https://doi.org/10.1177/09567976211024535* **32**, 1169–1178 (2021).
14. Rathje, S., Roozenbeek, J., Steenbuch Traberg, C., Van Bavel, J. J. & van der Linden, S. Letter to the Editors of Psychological Science: Meta-Analysis Reveals that AccuracyNudges Have Little to No Effect for U.S. Conservatives: Regarding Pennycook et al. (2020). *Psychological Science* (2022) doi:10.25384/SAGE.12594110.V2.
15. Guess, A., Nagler, J. & Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* **5**, eaau4586 (2019).
16. Pereira, A., Harris, E. A. & Bavel, J. J. V. Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. (2018) doi:10.31234/OSF.IO/7VC5D.
17. Sternisko, A., Cichocka, A., Cislak, A. & Bavel, J. J. V. National Narcissism predicts the Belief in and the Dissemination of Conspiracy Theories During the COVID-19 Pandemic: Evidence From 56 Countries: *https://doi.org/10.1177/01461672211054947* 014616722110549 (2021) doi:10.1177/01461672211054947.
18. Van Bavel, J. J. & Pereira, A. The Partisan Brain: An Identity-Based Model of Political Belief. *Trends in Cognitive Sciences* **22**, 213–224 (2018).
19. Cunningham, W. A. & Zelazo, P. D. Attitudes and evaluations : a social cognitive neuroscience perspective. **11**, (2007).
20. Rangel, A., Camerer, C. & Montague, P. R. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience 2008 9:7* **9**, 545–556 (2008).
21. Atran, S. & Ginges, J. Devoted actors and the moral foundations of intractable intergroup conflict. in *The moral brain: A multidisciplinary perspective* (eds. Decety, J. & Wheatley, T.) 69–85 (Boston Review, 2015).
22. Swann, W. B., Gómez, A., Seyle, D. C., Morales, J. F. & Huici, C. Identity fusion: the interplay of personal and social identities in extreme group behavior. *Journal of personality and social psychology* **96**, 995–1011 (2009).
23. Hogg, M. A. & Reid, S. A. Social identity, self-categorization, and the communication of group norms. *Communication Theory* **16**, 7–30 (2006).
24. Hogg, M. A., CooperShaw, L. & Holzworth, D. W. Group Prototypically and Depersonalized Attraction in Small Interactive Groups. *Personality and Social Psychology Bulletin* **19**, 452–465 (1993).
25. Baron, J. & Spranca, M. Protected Values. *Organizational Behavior and Human Decision Processes* **70**, 1–16 (1997).
26. Tetlock, P. E. Thinking the unthinkable: sacred values and taboo cognitions. *Trends in Cognitive Sciences* **7**, 320–324 (2003).
27. Atran, S. The devoted actor: Unconditional commitment and intractable conflict across cultures. *Current Anthropology* **57**, S192–S203 (2016).
28. Sachdeva, S. & Medin, D. Group identity salience in sacred value based cultural conflict: an examination of the Hindu-Muslim identities in the Kashmir and Babri mosque issues. *In Proceedings of the 31st Annual Conference of the Cognitive Science Society* 3111–3114 (2009).
29. Garrett, R. K. & Bond, R. M. Conservatives' susceptibility to political misperceptions. *Science Advances* **7**, eabf1234 (2021).
30. Weissman, D. H., Perkins, A. S. & Woldorff, M. G. Cognitive control in social situations: A role for the dorsolateral prefrontal cortex. *NeuroImage* **40**, 955–962 (2008).
31. Cohen, M. X. & Ranganath, C. Reinforcement Learning Signals Predict Future Decisions. **27**, 371–378 (2007).

32. Sheikh, H., Ginges, J. & Atran, S. Sacred values in the Israeli–Palestinian conflict: resistance to social influence, temporal discounting, and exit strategies. *Annals of the New York Academy of Sciences* **1299**, 11–24 (2013).
33. Yeo, B. T. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* **106**, 1125–65 (2011).
34. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data. *Frontiers in Neuroinformatics* (2011).
35. Kunst, J. R., Dovidio, J. F. & Thomsen, L. Fusion with political leaders predicts willingness to persecute immigrants and political opponents. *Nature Human Behaviour 2019 3:11* **3**, 1180–1189 (2019).
36. Nyhan, B. & Reifler, J. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior 2010 32:2* **32**, 303–330 (2010).
37. Van Bavel, J. J. *et al.* Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing. *Social Issues and Policy Review* **15**, 84–113 (2021).
38. Tajfel, H. & Turner, J. C. The Social Identity Theory of Intergroup Behavior. in *Psychology of Intergroup Relations* (eds. Worchel, S. & Austin, W. G.) 7–24 (Brooks Cole Publishing, 1986).
39. Kahan, D. M. Ideology, Motivated Reasoning, and Cognitive Reflection: An Experimental Study. *SSRN Electronic Journal* (2012) doi:10.2139/SSRN.2182588.
40. Kahan, D. M. Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition. *SSRN Electronic Journal* (2017) doi:10.2139/SSRN.2973067.
41. Dehghani, M. *et al.* Sacred values and conflict over Iran's nuclear program. *Judgment and Decision Making* **5**, 540–546 (2010).
42. Sheikh, H., Gómez, Á. & Atran, S. Empirical Evidence for the Devoted Actor Model. *Current Anthropology* **57**, S204–S209 (2016).
43. Abu-Akel, A. & Shamay-Tsoory, S. Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* **49**, 2971–2984 (2011).
44. Corradi-Dell'Acqua, C., Hofstetter, C. & Vuilleumier, P. Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Social Cognitive and Affective Neuroscience* **9**, 1175–1184 (2014).
45. Mitchell, J. P., Banaji, M. R. & Macrae, C. N. The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience* **17**, 1306–1315 (2005).
46. Harris, E. A. & Van Bavel, J. J. Preregistered Replication of "Feeling Superior Is a Bipartisan Issue: Extremity (Not Direction) of Political Views Predicts Perceived Belief Superiority". *Psychological Science* **32**, 451–458 (2021).
47. Kleiner, M. *et al.* What's new in psychtoolbox-3. *Perception* **36**, 1–16 (2007).
48. Green, P., MacLeod, C. & Alday, P. *Package 'simr' Type Package Title Power Analysis for Generalised Linear Mixed Models by Simulation*. (2016).
49. Hamid, N. *et al.* Neuroimaging 'will to fight' for sacred values: An empirical case study with supporters of an Al Qaeda associate. *Royal Society Open Science* **6**, (2019).
50. Weible, J. L. & Zimmerman, H. T. Science curiosity in learning environments: developing an attitudinal scale for research in schools, homes, museums, and the community. *International Journal of Science Education* **38**, 1235–1255 (2016).
51. Alfano, M. *et al.* Development and validation of a multi-dimensional measure of intellectual humility. *PLOS ONE* **12**, e0182950 (2017).
52. Simons, M. & Meeus, W. *Measuring Media Literacy for Media Education: Development of a Questionnaire for Teachers' Competencies*. *Sas / Journal of Media Literacy Education* vol. 9 99–115 www.jmle.org (2017).
53. Frederick, S. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* **19**, 25–42 (2005).
54. Singmann, H., Bolker, B., Westfall, J. & Aust, F. *afex: Analysis of Factorial Experiments. R package version 0.16-1.* (2016).

Acknowledgments

We want to thank all participants who took part in this study, especially those who participated in the fMRI study. We thank Marta Cazorla for support in developing materials, Luis Marcos for advice on the neuroimaging analysis, and Georgina Benet for assistance in the neuroimaging sessions.

Funding

This research has been funded by the European Commission, within the Enhanced European Innovation Council (EIC) Research and Innovation (RIA) action "FET Proactive - Boosting emerging technologies" (FETPROACT-EIC-05-2019).

Author contributions

J.V.B., O.V. and C.P. conceived the online and neuroimaging studies. J.V.B. directed the online experiments and managed project communication and O.V. directed the neuroimaging study. C.P. developed the online experiments and the neuroimaging study. W.J.B. and E.A.H. participated in developing the online studies and C.S. participated in data collection for the neuroimaging study. C.P. conducted the data analysis and initiated manuscript preparation. All authors edited and approved the final version of the manuscript.

Competing interests

Authors declare that they have no competing interests.

Data and code availability

The data and code employed in the analyses are available following this link (https://osf.io/twr6b/). Due to privacy concerns, the neuroimaging data can be obtained from the authors upon request with an approved IRB (we will send a password protected link to scientists with ethical approval).

Figures and Tables

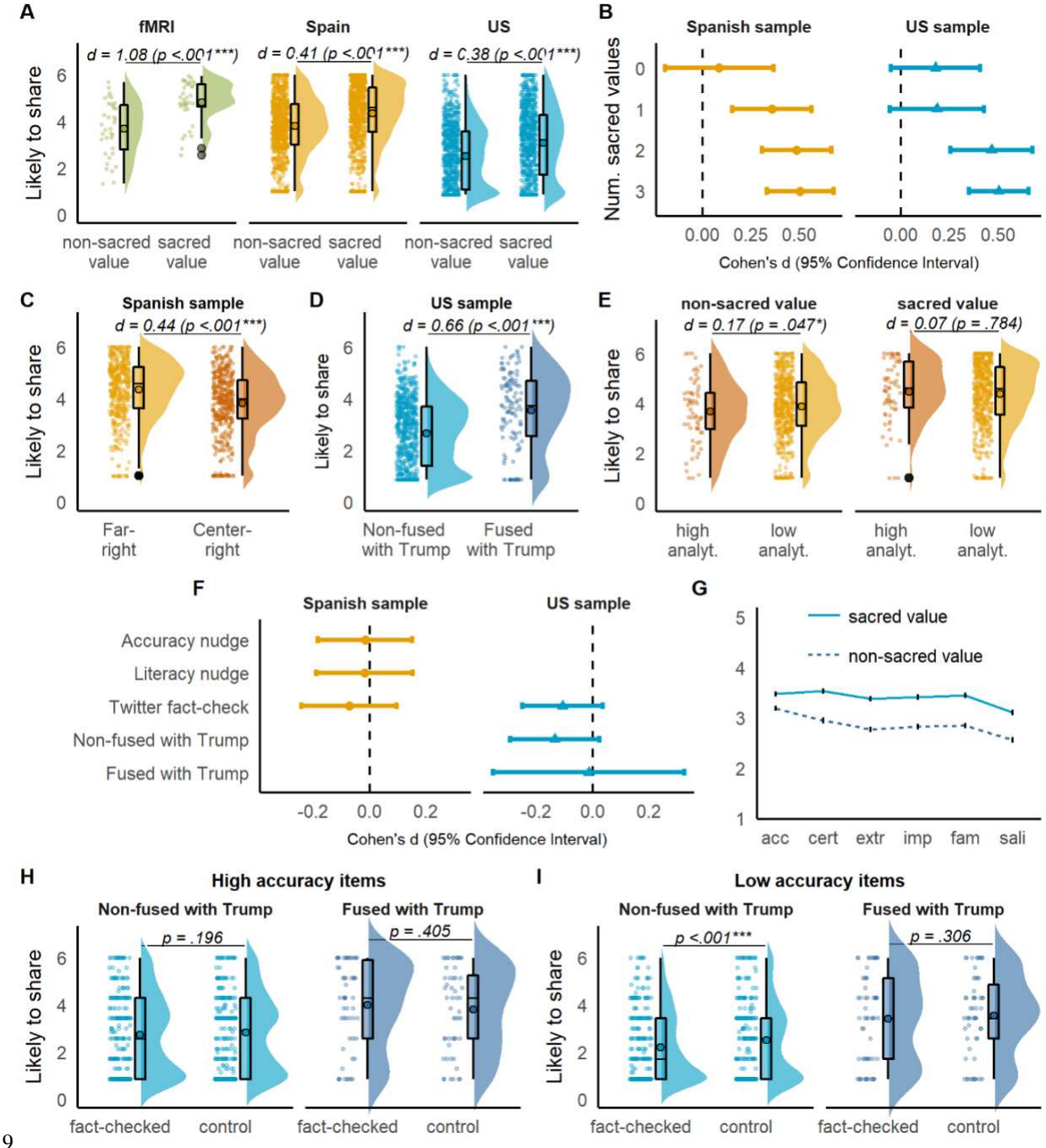

Fig. 1. Behavioral results: The effect of sacred vs. non-sacred values on likelihood of sharing was observed across the three samples **(A)** and was more prominent for participants with a higher number of sacred values in both the US and the Spanish sample **(B),** for conservatives fused vs. non-fused with Trump in the US **(C)** and for far-right vs. center-right voters in Spain **(D).** Analytical thinking

was associated with reduced sharing of misinformation about non-sacred values but was unrelated to sharing misinformation about sacred values in the Spanish survey sample **(E)** (the cognitive reflection test was not included in the US survey). Fact-checks were largely ineffective, with no effect of type of fact-check for fact-checked compared to control trials in the Spanish sample **(F left)** and a small effect of the Twitter fact-check in the experimental block (vs. baseline) in fact-checked participants compared to control participants in the US sample **(F right).** Fact-checks were even less effective in conservatives fused with Trump **(F right bottom)**. Boxplot bounds represent the interquartile range and the circle within the box represents the means.

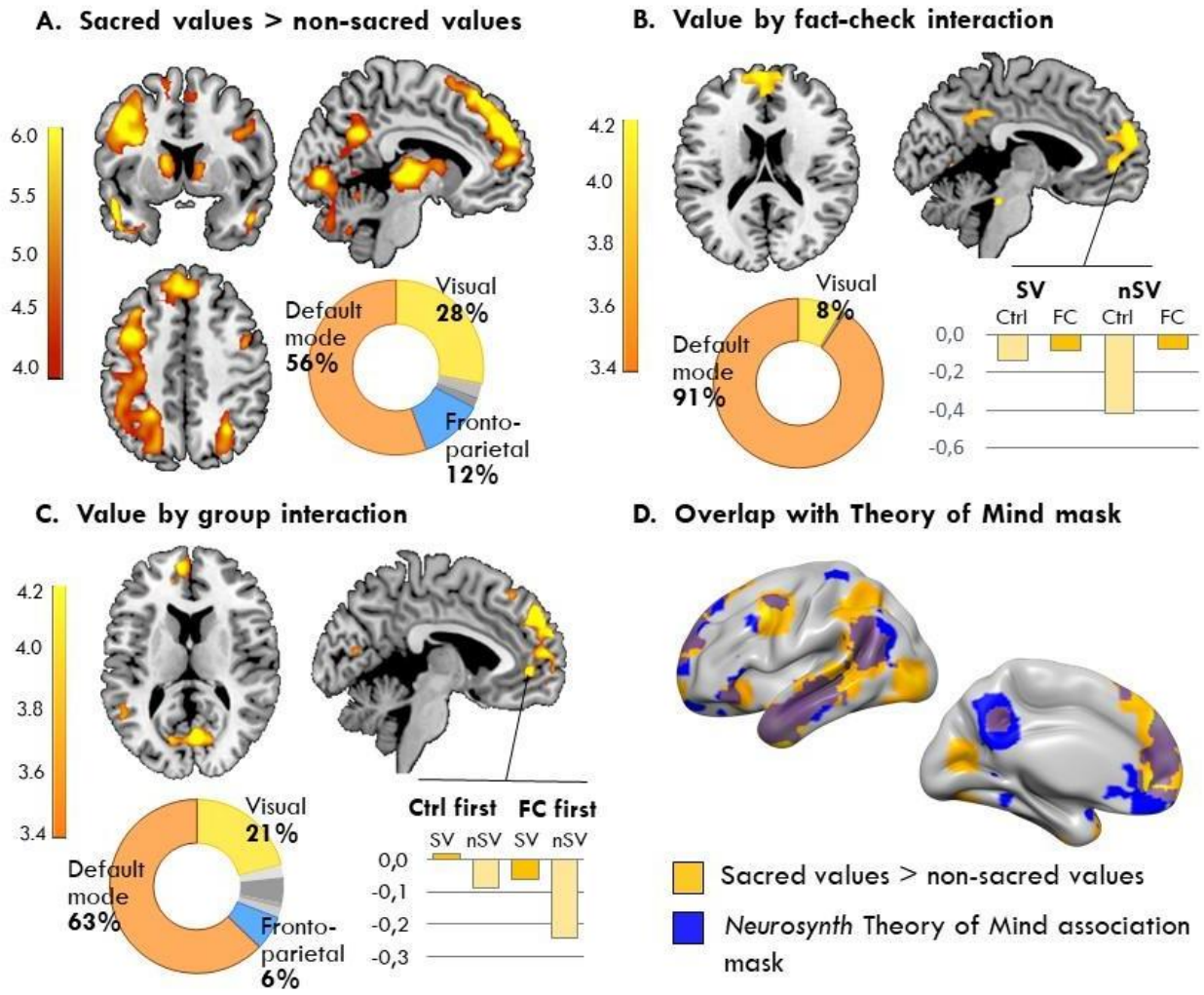

Fig. 2. Neuroimaging results. Social media posts with sacred values yielded much greater brain activation compared to posts with non-sacred values, especially in default mode regions **(A)**. A similar brain activation pattern was found for fact-checked trials compared to control trials, an increase that was more prominent for non-sacred value trials, which had notably lower activation in control trials than sacred value trials **(B)**. Participants who started with fact-checked trials exhibited lower activation in these brain areas as well, suggesting a reduced neural effect of fact-checks in participants who were familiarized with fact-checks from the beginning **(C)**. Comparison with a *Neurosynth* functional mask suggests that the activity detected in the sacred vs. non-sacred value contrast largely overlaps with functional activations associated with *theory of mind* studies. *SV = sacred values, nSV = non-sacred values, Ctr = control trials, FC = fact-checked trials.*

954 **Table 1. Demographic information of the study samples in Spain and the US.**

	Spanish sample (N = 812)		**US sample (N = 797)**		**fMRI ıple**
	Center-oters	*Far-right rs*	*Republican s Not fused rump*	*Republican s fused with mp*	*Spanish t voters*
N	404	408	675	122	36
% Women	60	48	46	47	33
Age (years)	45.6(15.6)	44.6(13.8)	41.7(13.7)	43.7(13.6)	23.1(4.8)
% Higher education	56	48	65	56	19
% Low household income ıth)	28*	32**	n/a	n/a	24***
Political orientation (1 = ght)	7.15(1.58)	7.55(1.84)	7.46(1.76)	8.5(1.85)	8.27(1.52)
Means of recruitment	YouGov Spain		Prolific		WhatsApp ups

955 *data not available for 63 participants; **data not available for 61 participants; ***data not available for 6
956 participants.

957

958 **Table 2. Behavioral results.** Effect of sacred values, political affiliation, and identity fusion on
959 willingness to share misinformation across the different study samples.

Effect		*M*	*SD*	M_{diff}	*95% CI*	*t*	*p*	*Cohen's d*	*95% CI*
(a) Sacred values									
Spanish sample	SV vs. non-SV			0.54	[0.48, 0.60]	18.80	< .001	0.41	[0.31, 0.51]
US sample	SV vs. non-SV			0.56	[0.50, 0.63]	16.03	< .001	0.38	[0.28, 0.48]
fMRI sample	SV vs. non-SV			1.14	[0.82, 1.47]	7.09	< .001	1.06	[0.58, 1.60]
(b) Political affiliation (Spain)									
Likelihood of sharing									
Far-right		4.35	1.55						
Center-right		3.81	1.61	0.54	[0.37, 0.71]	6.28	< .001	0.41	[0.31, 0.51]
Likelihood of sharing sacred values									
Interaction: Political affiliation x sacred values				0.24	[0.14, 0.35]	4.14	< .001		
Far-right	SV vs. non-SV			0.66	[0.58, 0.74]	16.38	< .001	0.52	[0.38, 0.66]
Center-right	SV vs. non-SV			0.42	[0.34, 0.50]	10.46	< .001	0.32	[0.19, 0.46]
Number of sacred values									
Far-right		2.09	0.95						
Center-right		1.62	1.03	0.46	[0.33, 0.60]	6.65	< .001	0.47	[0.33, 0.61]
(c) Fusion with far-right (Spain)									
Likelihood of sharing									
Fused		5.13	1.23						
Non-fused		4.21	1.57	0.92	[0.61, 1.24]	5.76	< .001	0.81	[0.52, 1.09]
Likelihood of sharing sacred values									
Interaction: Identity fusion x sacred values				0.08	[-0.15, 0.32]	0.68	*n.s.*(.50)		
Fused	SV vs. non-SV			0.59	[0.37, 0.81]	5.31	< .001	0.62	[0.25, 0.99]
Non-fused	SV vs. non-SV			0.67	[0.58, 0.76]	14.55	< .001	0.53	[0.38, 0.68]
Number of sacred values									
Fused		2.31	0.93						
Non-fused		2.04	0.95	0.27	[0.01, 0.53]	2.03	.043	0.28	[0.01, 0.56]
(d) Fusion with center-right (Spain)									
Likelihood of sharing									
Fused		4.48	1.58						
Non-fused		3.74	1.60	0.74	[0.31, 1.17]	3.41	< .001	0.60	[0.25, 0.95]
Likelihood of sharing sacred values									
Interaction: Identity fusion x sacred values				0.03	[-0.23, 0.29]	0.23	*n.s.*(.82)		
Fused	SV vs. non-SV			0.39	[0.14, 0.65]	3.05	.002	0.34	[-0.14, 0.82]
Non-fused	SV vs. non-SV			0.42	[0.35, 0.50]	10.71	< .001	0.32	[0.18, 0.47]
Number of sacred values									
Fused		1.83	1.07						

Non-fused		1.60	1.03	0.23	[-0.13, 0.59]	1.24	*n.s.*(.22)	0.22	[-0.13, 0.57]
Fusion with Republican party, excluding those fused with Trump (US)									
Likelihood of sharing									
Fused		3.73	1.87						
Non-fused		2.63	1.72	1.25	[0.89, 1.61]	6.77	< .001	0.99	[0.69, 1.28]
Likelihood of sharing sacred values									
Interaction: Identity fusion x sacred values				0.15	[0.06, 0.34]	1.08	*n.s.*(.28)		
Fused	SV vs. non-SV			0.37	[0.11, 0.64]	2.78	.006	0.30	[-0.10, 0.69]
Non-fused	SV vs. non-SV			0.52	[0.45, 0.60]	13.69	< .001	0.38	[0.27, 0.50]
Number of sacred values									
Fused		1.92	1.13						
Non-fused		1.67	1.20	0.24	[-0.08, 0.57]	1.46	*n.s.*(0.14)	0.21	[-0.07, 0.50]
(e) Fusion with Trump (US)									
Likelihood of sharing									
Fused		3.56	1.97						
Non-fused		2.64	1.71	0.89	[0.63, 1.15]	6.73	< .001	0.66	[0.47, 0.86]
Likelihood of sharing sacred values									
Interaction: Identity fusion x sacred values				0.32	[0.14, 0.52]	3.43	< .001		
Fused	SV vs. non-SV			0.84	[0.67, 1.02]	9.47	< .001	0.52	[0.27, 0.78]
Non-fused	SV vs. non-SV			0.51	[0.44, 0.59]	13.51	< .001	0.37	[0.26, 0.47]
Number of sacred values									
Fused		2.09	1.19						
Non-fused		1.85	1.14	0.24	[0.02, 0.46]	2.11	.035	0.21	[0.01, 0.40]

963 **Table 3. Neuroimaging results.** Neural activity resulting from (a) the contrast between sacred and
964 non-sacred values (*results of this contrast were reported at a more conservative threshold (0.05 FWE
965 peak-level) due to generalized activation across the brain at lower thresholds), (b) the interaction
966 contrast between sacred values and fact-checks, and (c) differences associated with sacred vs non-
967 sacred values in participants who started with control trials and those who started with fact-checked
968 trials.

			MNI coordinates			
Regional label	k	t(33)	x	y	z	P-FWEc
(a) ***Sacred > non-sacred values****						
L middle temporal gyrus	714	8.96	-52	-30	0	< .001
R middle occipital gyrus	966	8.67	36	-78	0	< .001
L inferior occipital gyrus	680	8.61	-30	-78	-4	< .001
L middle temporal gyrus	598	7.84	-50	-2	-20	< .001
L superior medial frontal gyrus	670	7.79	-4	48	38	< .001
L middle frontal gyrus	481	7.69	-36	12	40	< .001
R middle temporal gyrus	54	7.44	50	-16	-14	< .001
L thalamus	283	6.86	-10	-16	0	< .001
L precuneus	142	6.71	-8	-56	30	< .001
L inferior frontal gyrus	54	6.56	-48	12	18	< .001
(b) ***Sacred > non-sacred values; fact-checked > control trials***						
L hippocampus	339	5.29	-16	-26	-12	.010
L cerebellum	419	5.28	-30	-80	-30	.004
R superior medial frontal gyrus	1284	5.01	12	60	24	<.001
R cerebellum	425	4.98	24	-78	-38	.003
Cerebellar vermis	273	4.41	6	-68	-10	.024
R precuneus	402	4.19	10	-52	30	.004
(c) ***Sacred > non-sacred values; control first > fact-checked first***						
L superior medial frontal gyrus	1257	6.98	-4	52	38	<.001
L inferior occipital gyrus	298	6.41	-32	-80	-2	.021
L middle temporal gyrus	2405	5.11	-52	-36	4	<.001
R Calcarine	658	4.96	6	-72	12	<.001
Cerebellar vermis	511	4.35	8	-52	-22	.002

969
970
971
972
973
974

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplMaterial.pdf