Citation:

Pretus C., Javeed A., Hughes D.R., Hackburg K., Tsakiris M., Vilarroya O., Van Bavel J. (in press) The *Misleading* count: An identity-based intervention to mitigate the spread of partisan misinformation. *Philosophical Transactions B.*

The Misleading count:

An identity-based intervention to counter partisan misinformation sharing

Clara Pretus[1,2*], Ali M. Javeed[3], Diána Hughes[3], Kobi Hackenburg[4], Manos Tsakiris[4,5], Oscar Vilarroya[6], Jay J. Van Bavel[3*]

[1]*Department of Psychobiology and Methodology of Health Sciences, Universitat Autònoma de Barcelona, Spain*
[2]*Center of Conflict Studies and Field Research, ARTIS International, St Michaels, MD.*
[3]*Department of Psychology and Center for Neural Science, New York University, New York, NY*
[4]*Centre for the Politics of Feelings, School of Advanced Study, University of London, London, UK*
[5]*Department of Psychology, Royal Holloway, University of London, London, UK*
[6]*Department of Psychiatry and Forensic Medicine, Universitat Autònoma de Barcelona, Spain*

Keywords: Misinformation; Social Media; Social Norms; Social Identity; Intervention

*Authors for correspondence (clara.pretus@uab.cat ; jay.vanbavel@nyu.edu).
†Present address: Department of Psychology and Center for Neural Science, New York University, New York, NY

Main Text

Summary

Interventions to counter misinformation are often less effective for polarizing content on social media platforms. We sought to overcome this limitation by testing an identity-based intervention, which aims to promote accuracy by incorporating normative cues directly into the social media user interface. Across three pre-registered experiments in the U.S. (N=1,709) and UK (N=804), we found that crowdsourcing accuracy judgments by adding a Misleading count (next to the Like count) reduced participants' reported likelihood to share inaccurate information about partisan issues by 25% (compared to a control condition). The Misleading count was also more effective when it reflected in-group norms (from fellow Democrats/Republicans) compared to the norms of general users, though this effect was absent in a less politically polarized context (UK). Moreover, the normative intervention was roughly 5 times as effective as another popular misinformation intervention (i.e., the accuracy nudge reduced sharing misinformation by 5%). Extreme partisanship did not undermine the effectiveness of the intervention. Our results suggest that identity-based interventions based on the science of social norms can be more effective than identity-neutral alternatives to counter partisan misinformation in politically polarized contexts (e.g., the U.S.).

Introduction

Online misinformation poses a substantial threat to democracy and public health. From the *infodemic* surrounding the Covid-19 pandemic (1, 2) to the election fraud disinformation campaign which led to the January 6th assault on the U.S. Capitol (3), misinformation appears to be a significant risk to public health

and democratic institutions. On Twitter, falsehoods spread significantly farther, faster, and deeper than true stories–and this was especially true for political and emotional stories (Vosoughi et al., 2018). Online misinformation drives user engagement (4), capturing 2.3% of web traffic and 14% of Facebook engagement according to recent estimates (5). As such, social media companies have few incentives to eliminate misinformation (6). Existing infrastructure for online content moderation has also proven unable to meet rapidly increasing demand: moderation is often outsourced to foreign workers who need to make split-second decisions on content that is highly dependent on local social and political contexts (7). Therefore, it is critical to create systemic changes to social media infrastructure that can effectively reduce misinformation sharing in a way that is scalable, context-sensitive, and effective among at-risk groups. In the current paper, we develop and evaluate an identity-based intervention for reducing online misinformation sharing and compare it to popular approaches to reduce misinformation.

One of the most popular moderation-free approaches to combating misinformation is "accuracy nudging": presenting users with visual or textual cues which remind them to be accurate. This approach is based on the idea that people largely share misinformation because they are inattentive or lack analytical thinking skills (8) and has received extensive empirical investigation (9). However, recent studies and a meta-analysis suggest that the effect of accuracy nudges may be relatively weak (10), especially among conservatives, Republicans, and far-right supporters (11, 12). Similarly, a recent meta-analysis found that most strategies for debunking misinformation were not very effective overall (d = .19), and were even less effective when the issue was politically polarized (Chan & Albarracín, 2023). As such, there is an urgent need to develop effective and scalable correction strategies for misinformation in the political domain that works across the political spectrum.

According to the Identity-Based Model of Political Belief (13), individuals are more likely to believe and share misinformation when their partisan motives outweigh accuracy concerns (14, Pereira et al., 2023). This helps explain why partisan misinformation may be more difficult to debunk–especially in polarized contexts. For instance, we recently found that partisans who were highly devoted to a political party were more likely to spread misinformation than centrist votes and were unresponsive to fact-checking (12). The fact that interventions to counter misinformation based on analytical thinking are relatively ineffective for political extremists and right-wing users is practically important since these populations contribute the most to the spread of misinformation, at least in the US (15–17). It also underscores the need to incorporate social identity and group norms in the design of misinformation interventions.

Online misinformation is usually embedded in an interactive social environment (i.e., social media platforms) with visible social engagement metrics (e.g., number of *Likes*), which have been found to increase people's vulnerability to misinformation (18). However, this also offers great potential for interventions based on social psychology (19). For instance, actual reporting of fake news (20) and willingness to correct misinformation online (21) have been associated with social norms (i.e., beliefs on what others do or deem desirable, see (22)) about these behaviors. The effect of social norms may be more complex when it comes to misinformation about polarizing issues (e.g., election fraud allegations) since beliefs and behavior are likely to be determined by the intergroup and intragroup dynamics of fellow partisans. In line with Social Identity Theory (23) and Self-Categorization Theory (24), opinions from the in-group tend to induce greater conformity than opinions from the out-group (25). Thus, in a polarized digital environment, people may be more likely to conform to in-group social norms than to social norms by general users.

Here, we propose that exposing individuals to normative accuracy judgments by their in-group (*versus* general others) may be helpful to counter partisan misinformation (e.g., misinformation that favors specific in-group partisan stances). Indeed, laypeople are relatively good at distinguishing low-quality news content (26), raising the possibility of crowdsourcing accuracy judgments. This norms-based approach could be particularly useful for misinformation on politically polarizing issues (e.g., attitudes towards immigration, and universal healthcare, see (27, 28)), which people are more likely than misinformation on non-polarizing issues (e.g., infrastructure, see (12)). Crowdsourcing only from the in-group may also contribute to correcting inaccurate perceptions of in-group norms over particular issues, which could help reduce misperceived polarization (29). Therefore, identity-based interventions that leverage normative cues to nudge people into being more accurate, may be an effective and scalable approach to moderating online misinformation.

Current research

In the present work, we tested the effect of normative accuracy judgments from the in-group to reduce sharing of partisan misinformation in three pre-registered online experiments with Democrats and Republicans in the U.S. (N = 1,709) and Labor and Conservative voters in the UK (N = 804). Although both contexts are politically polarized, a cross-country analysis found higher levels of affective polarization in the U.S. compared to the UK (Boxell et al., 2022). We asked participants how likely they would be to share a series of simulated social media posts composed by different in-group political leaders (e.g., Bernie Sanders for Democrats) that contained inaccurate information relevant to politically polarizing issues (e.g., immigration, homelessness). The intervention consisted of adding a *Misleading* count next to the *Like* count. Half of the participants were told the *Misleading* count reflected in-group norms, i.e., the number of fellow Democrats/Republicans who had tagged the post as misleading (identity-relevant condition). The other half were told the *Misleading* count reflected the norms of general users (identity-neutral condition). We compared this intervention to widely used interventions to counter misinformation, including (a) the official Twitter tag, a warning that precludes social media users from further sharing the posts ("This Tweet can't be replied to, shared or liked"), and (b) an accuracy nudge adapted from Pennycook and Rand (30) ("To the best of your knowledge is the above statement accurate?"). This allowed us to test the relative efficacy of different popular interventions against the identity-based intervention.

We predicted that people would report a lower likelihood of sharing social media posts in response to seeing the *Misleading* count compared to no count (*H1*). In Studies 2 and 3, we also expected the *Misleading* count to be more effective in reducing sharing whenever the count was 80% compared to 20% of the *Like* count (*H2*). Since politically polarizing issues typically involve absolutist stances over moral issues, which are particularly resistant to trade-offs and social influence (31), we expected a reduced effect of the *Misleading* count when the posts were relevant to polarizing (vs. non-polarizing) issues in Experiment 1 (*H3* in Exp 1). We also expected the *Misleading* count to be similarly effective in reducing sharing of social media posts among Democrats and Republicans in the U.S., and among Labour and Conservative voters in the UK (*H3* in Exp 2 and 3). Finally, because people are responsive to in-group norms specifically (Hogg & Seid, 2006), we expected the Misleading count to be more effective when it reflected in-group norms compared to general users' norms (*H4*).

Methods

The data and code employed in the analyses are available at https://osf.io/dmxbt/. The pre-registrations for the three studies can be found at https://osf.io/xng3h (Experiment 1), https://osf.io/nmwvs (Experiment 2), and https://osf.io/m9hg3 (Experiment 3).

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. 5820).

Participants. We recruited 401 Democratic and 402 Republican voters in the US for Experiment 1, 402 Labour voters and 402 Conservative voters in the UK for Experiment 2, and 453 Democratic and 452 Republican voters in the US for Experiment 3 by means of an online panel (Prolific). Inclusion criteria included being 18 or older and having voted for the relevant political party in the two previous presidential elections (see demographic information in Table S1 and power analysis in Supplementary methods, Participants).

Materials. The posts were designed to look like Tweets and contained inaccurate information about a series of political issues that we expected would be engaging to participants. The use of artificial rather than real misinformation allowed us to control for both content and grammatical structure. The posts were tested for perceived accuracy, salience, familiarity, and importance in a series of pilot studies with independent samples matched for country of residence and political affiliation (see Supplementary methods, Materials, and Table S2). The pilot studies also confirmed that the information contained in the posts was neither too plausible nor too implausible to avoid ceiling and floor effects in participants' likelihood of sharing (see Table S2). Participants were exposed to information aligned with their political affiliation (e.g., in favor of universal healthcare for Democrats).

In Experiment 1, half of the social media posts included information about politically polarizing issues (e.g., immigration, universal healthcare), and the other half included non-polarizing issues (e.g., infrastructure). In Experiments 2 and 3, all the social media posts conveyed information about polarizing issues (e.g.,

immigration, universal healthcare). As expected, a larger proportion of participants held absolutist stances (resistant to economic trade-offs) over issues that we proposed as polarizing as compared to issues that we proposed as non-polarizing. Whether participants held absolutist stances over each issue was assessed in the same survey (see Supplementary Materials and Table S3).

Procedure. We launched surveys asking Democrats and Republicans in the U.S. (Experiments 1 and 3) and Labour and Conservative voters in the UK (Experiment 2) to rate the likelihood of sharing a series of social media posts composed by different political leaders of the party they voted for in the last elections. We tested different variations of an identity-based intervention: we included a *Misleading* count next to the *Like* count, which we told participants reflected in-group norms, i.e., the number of fellow Democrats/Republicans who had tagged the post as misleading. The *Misleading* count was always lower than the *Like* count, as expected for highly partisan content. Social media posts with and without interventions were presented in a randomized order.

In Experiment 1 (N = 803), half of the social media posts contained a *Misleading* count which was 30% of the *Like* count (see Fig. 1a) and the other half did not contain any intervention (control condition).

In Experiment 2 (N = 804), 25% of the social media posts contained a low *Misleading* count (20% of the Like count), 25% contained a high *Misleading* count (80% of the Like count), 25% contained an official Twitter Misleading tag (see Fig 1b), and 25% did not include any intervention (control condition). Because the official Twitter Misleading tag prevents participants from sharing the post, we asked participants how likely they would be to share the post through other means (e.g., taking a screenshot). Since sharing messages in this condition requires extra effort, the dependent variable (sharing intentions) is psychologically different in the official Twitter tag compared to the Misleading count conditions. The value of including an established intervention against misinformation (the official Twitter tag) lies in the comparison of final outcomes (whether the message is likely to be shared or not, independently of the means), which is relevant for potential implementations of the *Misleading* count.

In Experiment 3 (N = 905), we used the same design as in Experiment 2 but instead of the official Twitter Misleading tag, we compared the high and low *Misleading* count to an accuracy nudge ("To the best of your knowledge, is the above statement accurate?", adapted from Pennycook & Rand, 2020). Of note, the accuracy nudge was included directly on the social media posts (see Fig. 1c) unlike in the original setting, where it was administered as a separate intervention at the beginning of the experiment (Pennycook & Rand, 2020). Thus, participants were exposed to the accuracy nudge more intensively than in the original setting. As such, the accuracy nudge effects might be stronger here than in the traditional implementation.

Across the three studies, all participants were exposed to all interventions and the control condition (within-subjects factor). All three studies included an additional between-subjects control condition so that half the sample was told that the *Misleading* count reflected general users' norms, i.e., the number of general users who had tagged the post as misleading ("tagged by anyone") instead of only fellow Democrats/Republicans ("tagged by in-group") (see Supplementary Methods for more details).

Results

Effect of the intervention. As predicted (*H1*), including a Misleading count next to the Like count (see Fig. 1a) reduced participants' likelihood of sharing misinformation compared to the no intervention control condition across all three studies ($p < .001$, Cohen's $d = 0.20$, see Table S4a and Fig. 2). In addition, participants were sensitive to the proportion of Misleadings compared to the number of Likes (*H2*) both in the U.S. (Experiment 2) and the UK (Experiment 3). Specifically, respondents reported a slightly lower likelihood of sharing when the Misleading count was 80% compared to 20% of the Like count in the U.S., $M_{diff} = -0.14$, 95% CI [-0.24, -0.04], z-score = -3.56, $p = .002$, $d = 0.08$, and in the UK, $M_{diff} = -0.16$, 95% CI [-0.28, -0.07], z-score = -4.43, $p = .001$, $d = 0.10$. In Experiment 2, the high Misleading count condition (80% of the Like count) was outperformed by the official Twitter Misleading tag, $M_{diff} = -0.17$, 95% CI [-0.26, -0.08], z-score = -4.61, $p < .001$, $d = 0.12$, which prevents Twitter users from further sharing the post ("This Tweet can't be replied to, shared or liked", see Fig. 1b). However, in Experiment 3, the high Misleading count worked better than the accuracy nudge (see Fig. 1c) in reducing participants' likelihood of sharing misinformation, $M_{diff} = -0.22$, 95% CI [-0.32, -0.12], z-score = -5.60, $p < .001$, $d = 0.13$. As such, the Misleading count appears to be a relatively effective strategy for reducing misinformation sharing.

In an explorative analysis looking at the dichotomized response variable (likely *versus* not likely to share) across the three experiments, the number of participants likely to share the social media posts (likelihood > 3) was reduced by around 25% in response to the *Misleading* count *versus* control in the in-group condition, M_{diff} = -0.26, *95% CI* [-0.35, -0.16], t(95) = -6.85, $p < .001$, as compared to a 12% reduction in the general users' condition *versus* control, M_{diff} = -0.12, *95% CI* [-0.22, -0.03], t(95) = -3.30, $p = .007$ (interaction effect: B = 0.13, *95% CI* [0.03, 0.23], t(94) = 2.51, $p = .014$). The same model revealed a 34% reduction in the number of participants likely to share the social media posts in response to the official Twitter tag *versus* control (B = -0.34, *95% CI* [-0.45, -0.23], t(98) = -6.13, $p < .001$) and a 5% reduction in response to the accuracy nudge *versus* control (B = -0.05, *95% CI* [-0.16, 0.06], t(98) = -0.91, $p = .366$).

To test the possibility of demand effects (i.e., to see if the "study would become quite obvious" over time), we tested if the initial effects (first 4 trials) were different than the overall effects in an exploratory analysis. Presumably, the effect would change over time if demand effects increasingly came into play. However, the results of this additional analysis revealed that the early effects were nearly identical to the overall effects (see Table S8). Thus, potential demand effects do not appear to have changed our results in any measurable way.

Extreme partisanship. Interventions to counter misinformation are often less effective when partisan incentives outweigh accuracy concerns, for instance, when misinformation is framed in terms of group-relevant politically polarizing issues, and for individuals who highly identify with the group (12). Thus, we tested if the *Misleading* count was also less effective in these conditions. In Experiment 1, we compared the effect of the intervention for misinformation relevant to politically polarizing issues compared to non-polarizing issues (as measured in our surveys, see the percentage of participants with absolutist stances over each issue in Table S3). Contrary to our expectations (*H3* in Exp 1), the *Misleading* count (*versus* control) was actually *more* effective for social media posts on polarizing issues (e.g., immigration, homelessness) than non-polarizing issues (interaction effect: B = -0.13, *95% CI* [-0.25, -0.002], t(2406) = -2.00, $p = .046$, see Fig. 2a, b).

In terms of identity fusion (i.e., visceral oneness with a group or leader, see (32)), we found no interaction effect between the intervention and identity fusion with the leader or with the political party ($p > 0.093$). If anything, there was a trend in the opposite direction in Experiment 1 (interaction effect: B = -0.16, *95% CI* [-0.36, -0.03], t(801) = -1.68, $p = .093$). Specifically, participants who reported feeling fused with the leader were more responsive to the *Misleading* count (versus control), M_{diff} = -0.44, *95% CI* [-0.62, -0.26], t(801) = -8.17, $p < .001$, d = 0.27, than non-fused participants, M_{diff} = -0.28, *95% CI* [-0.35, -0.21], t(801) = -8.17, $p < .001$, d = 0.19. Therefore, extreme partisanship measured as both identity fusion with leaders and with political parties did not undermine the effectiveness of the intervention.

Political affiliation. In contrast to our pre-registered hypothesis (*H3* in Exp 2 and 3), Democrats and Labour voters were generally more responsive to the intervention than Republicans and Conservatives, respectively (see Table S4c and Fig. 2a, c, e). This effect was less clear in Experiment 1, where the *Misleading* count (*versus* control) was only marginally better ($p = .072$) at reducing the likelihood of sharing misinformation among Democrats, M_{diff} = -0.36, *95% CI* [-0.47, -0.27], t(801) = -7.88, $p < .001$, d = 0.23, compared to Republicans, M_{diff} = -0.24, *95% CI* [-0.33, -0.15], t(801) = -5.34, $p < .001$, d = 0.17. In Experiment 2, group differences were most apparent in the high misleading condition, where Labour voters in the UK reduced their likelihood of sharing in response to the intervention (*versus* control) to a greater extent, M_{diff} = -0.42, *95% CI* [-0.56, -0.29], z-score = -8.20, $p < .001$, d = 0.29, than Conservative voters, M_{diff} = -0.14, *95% CI* [-0.28, -0.02], z-score = -2.90, $p = .020$, d = 0.11. However, U.S. Republicans and UK Conservatives were overall less likely to share the social media posts that were presented to them both in Experiment 1, M_{diff} = -0.31, *95% CI* [-0.48, -0.14], t(801) = -3.58, $p < .001$, d = 0.21 (see Fig. 3a), and in Experiment 2, M_{diff} = -0.98, *95% CI* [-1.16, -0.81], t(802) = -10.89, $p < .001$, d = 0.68 (see Fig. 3d). Thus, it could be that the reduced effect of the intervention in these samples was due to floor effects in likelihood of sharing.

Therefore, in Experiment 3, we made sure the stimuli for Republicans were matched with those presented to Democrats in perceived accuracy, attitude strength, familiarity, and salience (see Table S2 and Fig. 3g). As a result, we found only a marginal difference in overall sharing between groups in Experiment 3, M_{diff} = -

0.18, *95% CI* [-0.38, 0.01], t(897) = -1.83, p = .067, d = 0.11, and if anything, the trend indicated higher sharing among Republicans (see Fig. 3h). Despite having similarly appealing social media posts for Democrats and Republicans in Experiment 3, the effect of the intervention (*versus* control) was still larger in Democrats than Republicans (see Table S4c and Fig. 2e), which is consistent with the accuracy nudge intervention (11).

Group norms. In line with our hypothesis (*H4*), the *Misleading* count was more effective when it reflected in-group norms (i.e., the number of fellow Republicans/Democrats who had tagged the post as misleading), as compared to the norms of general users (see Table S4b and Fig. 2b, d, f). Specifically, in Experiment 1 (intervention x in-group interaction: B = -0.16, *95% CI* [-0.29, -0.04], t(801) = -2.57, p = .01), the *Misleading* count (*versus* control) was associated with a steeper reduction in the likelihood of sharing in the in-group condition, M_{diff} = -0.38, *95% CI* [-0.47, -0.29], t(810) = -8.40, p < .001, d = 0.26, compared to the general users' condition, i.e., tagged by anyone, M_{diff} = -0.22, *95% CI* [-0.31, -0.13], t(801) = -4.88, p < .001, d = 0.14. In Experiment 3, the effect of the in-group was particularly important for the low misleading count condition (intervention x in-group interaction: B = -0.17, *95% CI* [-0.34, -0.04], t(2694) = -2.25, p = .024), such that the low *Misleading* count (20% of the Likes) was associated with reduced misinformation sharing compared to control only in the in-group condition, M_{diff} = -0.27, *95% CI* [-0.41, -0.13], z-score = -4.91, p < .001, d = 0.16, but not in the general users' condition, M_{diff} = -0.09, *95% CI* [-0.24, 0.05], z-score = -1.72, p = .312, d = 0.06. Similarly, the high *Misleading* count (80% of the Likes) *versus* control was more effective in the in-group condition, M_{diff} = -0.44, *95% CI* [-0.58, -0.30], z-score = -7.97, p < .001, d = 0.26, than the general users' condition, M_{diff} = -0.20, *95% CI* [-0.34, -0.06], z-score = -3.71, p = .001, d = 0.12 (intervention x in-group interaction: B = -0.23, *95% CI* [-0.38, -0.08], t(2694) = -3,01, p = .003).

However, the effect of the in-group was only observed in the U.S. sample (Experiments 1 and 3) and was not found in the UK sample (Experiment 2, p > 0.194, see Table S4b), which was less polarized in terms of political orientation (see Fig. 3f) than the U.S. samples (see Fig. 3c and 3i). Thus, incorporating the in-group dimension may be more effective in highly polarized U.S. voters (see political orientation distribution in Fig. 3C and 3I) and but less so among less polarized UK voters (see political orientation distribution in Fig. 3F).

Engagement. The social media posts used in Experiment 1 were designed to have higher social engagement metrics (from 782 to 16900 *Misleadings*) than those in Experiments 2 and 3 (from 62 to 199 *Misleadings*, see all items in Table S5, S6, and S7). In an explorative analysis, we found that participants were less likely to share misinformation the higher the absolute count of *Misleadings* was (Experiment 1: B = -0.00002, *95% CI* [-0.00003, -0.00001], t(9721) = -3.51, p < .001; Experiment 2: B = -0.0006, *95% CI* [-0.0008, -0.0004], t(2432) = -7.36, p < .001; Experiment 3: B = -0.0005, *95% CI* [-0.0007, -0.0004], t(3036) = -6.33, p < .001). Thus, more *Misleadings* were associated with less likelihood of sharing, and the cumulative effect of *Misleadings* in high engagement Tweets was enough to counteract the effect of the *Misleading* to *Likes* ratio. If this ratio was all that mattered, the *Misleading* count in Experiment 1 (30% of the *Like* count) would be less effective in reducing sharing than the high *Misleading* count in Experiments 2 and 3 (80% of the Like count). However, in an explorative analysis combining the three data sets, we found the *Misleading* count in Experiment 1 to be similarly effective in reducing sharing than the high *Misleading* count in Experiments 2 and 3, that is, the interaction term between Experiment and intervention was non-significant (Experiment 1 vs Experiment 2: B = -0.08, *95% CI* [-0.19, 0.03], t(2504) = 1.30, p = .195; Experiment 1 vs Experiment 3: B = -0.05, *95% CI* [-0.16, -0.06], t(2504) = 0.850, p = .395). Thus, participants were not only responsive to the *Like* to *Misleading* count ratio but also to the absolute number of *Misleadings*.

Discussion

We tested the effect of an identity-based misinformation intervention by adding a *Misleading* count next to the *Like* count on social media posts to reduce sharing of partisan misinformation in the U.S. and the UK. Across three experiments, the number of people who reported they would be likely to share the social media posts dropped by 25% in response to the in-group *Misleading* count (vs. control) as compared to 5% in response to an adapted version of the accuracy nudge. The *Misleading* count was also more effective when it reflected in-group norms (fellow Democrats/ Republicans) compared to the norms of general users and when it was relatively higher compared to the *Like* count. The effect of the in-group was not found in the UK sample, which was less politically polarized. Moreover, extreme partisanship, measured as both identity

fusion with the political party and identity fusion with leaders, did not undermine the effectiveness of the intervention. These results provide initial evidence that identity-based interventions may be more effective than identity-neutral alternatives for addressing partisan misinformation in polarized contexts.

While completely preventing users from engaging with misinformation – the official Twitter tag condition - was most effective, this strategy relies heavily on moderators and is better suited for unequivocally false rather than misleading content. The *Misleading* count provides an additional and easily scalable layer against misinformation where social media users are able to regulate online content themselves. It had larger cumulative effects (25% fewer people sharing across experiments) than identity-neutral alternatives such as the accuracy nudge (5% fewer people sharing), and it was effective for extreme partisans. The *Misleading* count allows social media users to update their perceived social norms about a given message, which can lead people to conform to these judgments and make behavioral adjustments (McDonald & Crandall, 2015). In contrast to a *Dislike* count, the *Misleading* count provides normative information on the accuracy of the message. Thus, its function is to convey information about the quality of a message rather than the level of agreement with a given statement.

Using normative cues seems to be more effective in polarized contexts (e.g., U.S. voters compared to UK voters) and for posts on polarizing issues (e.g., immigration) compared to non-polarizing issues (e.g., infrastructure). These findings suggest that polarized contexts offer either greater incentives to conform to in-group norms or greater disincentives not to conform to them–which is consistent with an identity-based approach to misinformation (12-14). As a result, people are more attuned to in-group (*versus* outgroup) norms in highly polarized contexts (33), especially when the issue at stake is fundamental to their status as group members. The effectiveness of in-group norms when group status is most salient (e.g., in polarized contexts and for posts on polarizing issues) also helps clarify why the *Misleading* count was effective even for extreme partisans who reported being fused with either political parties or leaders. Because fused individuals are more driven to match in-group norms (Pretus & Vilarroya, 2022), the *Misleading* count and other norm-based interventions appear to be particularly compelling for extreme partisans (see also Hamid, Pretus, et al., 2019).

Our design offers novel data on the motivations underlying individuals' responses to the *Misleading* count. Partisans could use the in-group norms to identify the most effective posts to promote their views (competitive effectiveness hypothesis). This mechanism could trigger a backfire effect, increasing people's likelihood of sharing posts with a relatively low *Misleading* count compared to the *Like* count, over what would have been expected without any intervention (control condition). However, we did not find support for this hypothesis: even relatively few *Misleadings* (vs. *Likes*) reduced participants' likelihood of sharing posts compared to the control condition. Conversely, partisans could use in-group norms as a trusted source to identify true information (civic-mindedness hypothesis). This would involve reductions in sharing in response to any number of *Misleadings* irrespective of *Likes*. In line with this hypothesis, we found an effect of the absolute number of *Misleadings* in reducing participants' likelihood of sharing posts. Future studies should explore these differences in greater depth.

While the inclusion of a *Misleading* button that people can click on is straightforward to implement and scale, incorporating the in-group condition is more challenging. One option would be to replace the in-group (fellow Democrats/Republicans) with "people you follow" or by AI-generated user subgroups (e.g., "people like you"). In this case, the *Misleading* count would only be altered by a particular subgroup, and each user would see a different *Misleading* count, preventing outgroup members from abusing the *Misleading* button to "attack" social media posts. This intervention relies on naturally occurring variation in how the in-group evaluates a particular social media post (some will *Like* it, others will tag it as misleading). While it is not clear how many in-group members will be willing to tag specific in-group misinformation as misleading, we found that just a few *Misleading tags* (e.g., 20% of the *Like* count) are enough to have a deterrent effect. Future research should assess if 20% is a realistic assumption and see if people are willing to tag in-group content as misleading. Prior research suggests that crowdsourcing accuracy judgments may be feasible and effective (26).

Similarly to other interventions such as the accuracy nudge (11, 12), the *Misleading* count is less effective for conservatives than liberals. This partisan difference could be related to perceptions of the existing norms within these political groups, labeling interventions as punitive and biased (35), or psychological differences in need for closure (36) and accuracy motivation (15). In the case of the *Misleading* count, this limitation

can be partially compensated with a higher total *Misleading* count (e.g., in Experiment 1). Thus, between-group differences in the effect of the intervention notably decrease for high-engagement Tweets with higher total *Misleading* counts.

The main limitation of the present study is that it is a series of controlled experiments with carefully curated social media posts. Although intentions to share are highly correlated with real-world sharing (Mosleh et al, 2020), more ecologically valid approaches are necessary to determine its effectiveness in a real-world social media setting. This is especially important for intervention studies that report small to medium effect sizes in samples of panel respondents such as the present study. In a more realistic setting, people would be exposed to a collection of both partisan and nonpartisan social media posts with accurate and inaccurate information and would be able to actually share the posts with others. Related to this, although our *likelihood of sharing* measure is widely employed to assess intentions to share social media posts (e.g., (26)), it could be that it overestimates people's actual likelihood to share posts in the real world. Future research could thus test the proposed intervention within a social media simulation or in field studies.

Moreover, we did not measure how perceived norms about the accuracy of each message changed before and after the intervention. Thus, we cannot directly test whether shifts in perceived social norms mediate the effect of the intervention on sharing intentions. Finally, the current research includes liberal and conservative voters in the US and the UK, limiting the generalizability of the findings to these populations.

Conclusions

Identity-based interventions which incorporate normative cues appear to be more effective than identity-neutral interventions to counter partisan misinformation among individuals in politically polarized contexts (e.g., U.S. voters). Particularly, pairing partisan misinformation with in-group accuracy judgments reduced misinformation sharing among partisans in the US and the UK. This strategy was effective even for extreme partisans who highly identified with their political leader. Thus, allowing social media users to publicly tag posts as misleading could contribute to stopping the spread of misinformation.

Acknowledgments

This research has been funded by the European Innovation Council (FETPROACT-EIC-05-2019). MT and KH were supported by the NOMIS Foundation.

Author Contributions

C.P., J.V.B., and O.V. conceptualized the studies. J.V.B. supervised the studies and C.P. conducted the investigation and the formal analysis. A.M.J., D.H., K.H., and M.T. developed the methodology. C.P. prepared the original draft. All authors reviewed and edited the final version of the manuscript.

Competing Interest Statement

The authors declare no competing interest.

References

1. S. van der Linden, Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine 2022 28:3* **28**, 460–467 (2022).

2. WHO, Infodemic (2021) (June 21, 2022).

3. C. Silverman, C. Timberg, J. Kao, J. B. Merrill, Facebook Hosted Surge of Misinformation and Insurrection Threats in Months Leading Up to Jan. 6 Attack, Records Show . *ProPublica | The Washington Post* (2022) (June 21, 2022).

4. L. Edelson, *et al.*, Far-right news sources on Facebook more engaging. *Cybersecurity for Democracy* (2021) (June 23, 2022).

5. S. Altay, R. K. Nielsen, R. Fletcher, Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media* **2**, 1–29 (2022).

6. H. Rahman, Why Are Social Media Platforms Still So Bad at Combating Misinformation? *Kellogg Insight* (2020) (June 21, 2022).

7. A. Arsht, D. Etcovitch, The Human Cost of Online Content Moderation. *Harv J Law Technol* (2018) (June 21, 2022).

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

9. G. Pennycook, D. G. Rand, Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications 2022 13:1* **13**, 1–12 (2022).

10. J. Roozenbeek, A. L. J. Freeman, S. van der Linden, How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020): *https://doi.org/10.1177/0956797621102453 5* **32**, 1169–1178 (2021).

11. S. Rathje, J. Roozenbeek, C. Steenbuch Traberg, J. J. van Bavel, S. van der Linden, 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). *Psychol Sci* (2022) https:/doi.org/10.25384/SAGE.12594110.V2 (January 28, 2022).

12. C. Pretus, *et al.*, The role of political devotion in sharing partisan misinformation (2022) https:/doi.org/10.31234/osf.io/7k9gx.

13. J. J. van Bavel, A. Pereira, The Partisan Brain: An Identity-Based Model of Political Belief. *Trends Cogn Sci* **22**, 213–224 (2018).

14. D. Borukhson, P. Lorenz-Spreen, M. Ragni, When Does an Individual Accept Misinformation? An Extended Investigation Through Cognitive Modeling. *Comput Brain Behav* **5**, 244–260 (2022).

15. R. K. Garrett, R. M. Bond, Conservatives' susceptibility to political misperceptions. *Sci Adv* **7**, eabf1234 (2021).

16. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election. *Science (1979)* **363**, 374–378 (2019).

17. A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Asian-Australas J Anim Sci* **32** (2019).

18. M. Avram, N. Micallef, S. Patil, F. Menczer, Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review* (2020) https:/doi.org/10.37016/mr-2020-033 (June 23, 2022).

19. J. J. van Bavel, *et al.*, Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing. *Soc Issues Policy Rev* **15**, 84–113 (2021).

20. H. Gimpel, S. Heger, C. Olenberger, L. Utz, The Effectiveness of Social Norms in Fighting Fake News on Social Media. *https://doi.org/10.1080/07421222.2021.1870389* **38**, 196–221 (2021).

21. A. Z. X. Koo, M. H. Su, S. Lee, S. Y. Ahn, H. Rojas, What Motivates People to Correct Misinformation? Examining the Effects of Third-person Perceptions and Perceived Norms. *https://doi.org/10.1080/08838151.2021.1903896* **65**, 111–134 (2021).

22. R. I. Mcdonald, C. S. Crandall, Social norms and social influence. *Curr Opin Behav Sci* **3**, 147–151 (2015).

23. H. Tajfel, J. C. Turner, "The Social Identity Theory of Intergroup Behavior" in *Psychology of Intergroup Relations*, S. Worchel, W. G. Austin, Eds. (Brooks Cole Publishing, 1986), pp. 7–24.

24. J. C. Turner, "Social categorization and the self-concept: A social cognitive theory of group behavior" in *Advances in Group Processes: Theory and Research*, E. J. Lawler, Ed. (JAI Press, 1985), pp. 77–121.

25. D. Abrams, M. Wetherell, S. Cochrane, M. A. Hogg, J. C. Turner, Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology* **29**, 97–119 (1990).

26. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc Natl Acad Sci U S A* **116**, 2521–2526 (2019).

27. J. Baron, M. Spranca, Protected Values. *Organ Behav Hum Decis Process* **70**, 1–16 (1997).

28. P. E. Tetlock, Thinking the unthinkable: sacred values and taboo cognitions. *Trends Cogn Sci* **7**, 320–324 (2003).

29. J. Lees, M. Cikara, Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B* **376** (2021).

30. G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, D. G. Rand, Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention: *Psychol Sci* **31**, 770–780 (2020).

31. H. Sheikh, J. Ginges, S. Atran, Sacred values in the Israeli–Palestinian conflict: resistance to social influence, temporal discounting, and exit strategies. *Ann N Y Acad Sci* **1299**, 11–24 (2013).

32. W. B. Swann, Á. Gómez, D. C. Seyle, J. F. Morales, C. Huici, Identity fusion: The interplay of personal and social identities in extreme group behavior. *J Pers Soc Psychol* **96**, 995–1011 (2009).

33. E. J. Finkel, *et al.*, Political sectarianism in America: A poisonous cocktail of othering, aversion, and moralization poses a threat to democracy. *Science (1979)* **370**, 533–536 (2020).

34. C. Kang, S. Frenkel, Republicans Accuse Twitter of Bias Against Conservatives. *The New York Times* (2018) (June 22, 2022).

35. E. Saltz, C. R. Leibowicz, C. Wardle, Encounters with Visual Misinformation and Labels across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. *Conference on Human Factors in Computing Systems -*

Proceedings (2021) https:/doi.org/10.1145/3411763.3451807 (June 23, 2022).

36. J. T. Jost, J. Glaser, A. W. Kruglanski, F. J. Sulloway, Political Conservatism as Motivated Social Cognition. *Psychol Bull* **129**, 339–375 (2003).

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

Chan, Mp.S., Albarracín, D. A meta-analysis of correction effects in science-relevant misinformation. Nat Hum Behav (2023). https://doi.org/10.1038/s41562-023-01623-8

Pereira, A., Harris, E., & Van Bavel, J. J. (2023). Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations*, *26*(1), 24-47.

McDonald, R. I., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences*, *3*, 147-151.

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication theory*, *16*(1), 7-30.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2022). Cross-country trends in affective polarization. *Review of Economics and Statistics*, 1-60.

Pretus, C., & Vilarroya, Ó. (2022). Social norms (not threat) mediate willingness to sacrifice in individuals fused with the nation: Insights from the COVID-19 pandemic. *European Journal of Social Psychology*, *52*(4), 772-781.

Hamid, N., Pretus, C., Atran, S., Crockett, M. J., Ginges, J., Sheikh, H., ... & Vilarroya, O. (2019). Neuroimaging 'will to fight'for sacred values: an empirical case study with supporters of an Al Qaeda associate. *Royal Society open science*, *6*(6), 181585.

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos one*, *15*(2), e0228882.

Figures

Figure 1. Employed interventions. Examples of the employed interventions including **(A)** the *Misleading* count condition (30% of the Like count) used in Exp. 1**, (B)** the official Twitter Misleading tag used in Exp. 2, and **(C)** the accuracy nudge used in Exp.3.

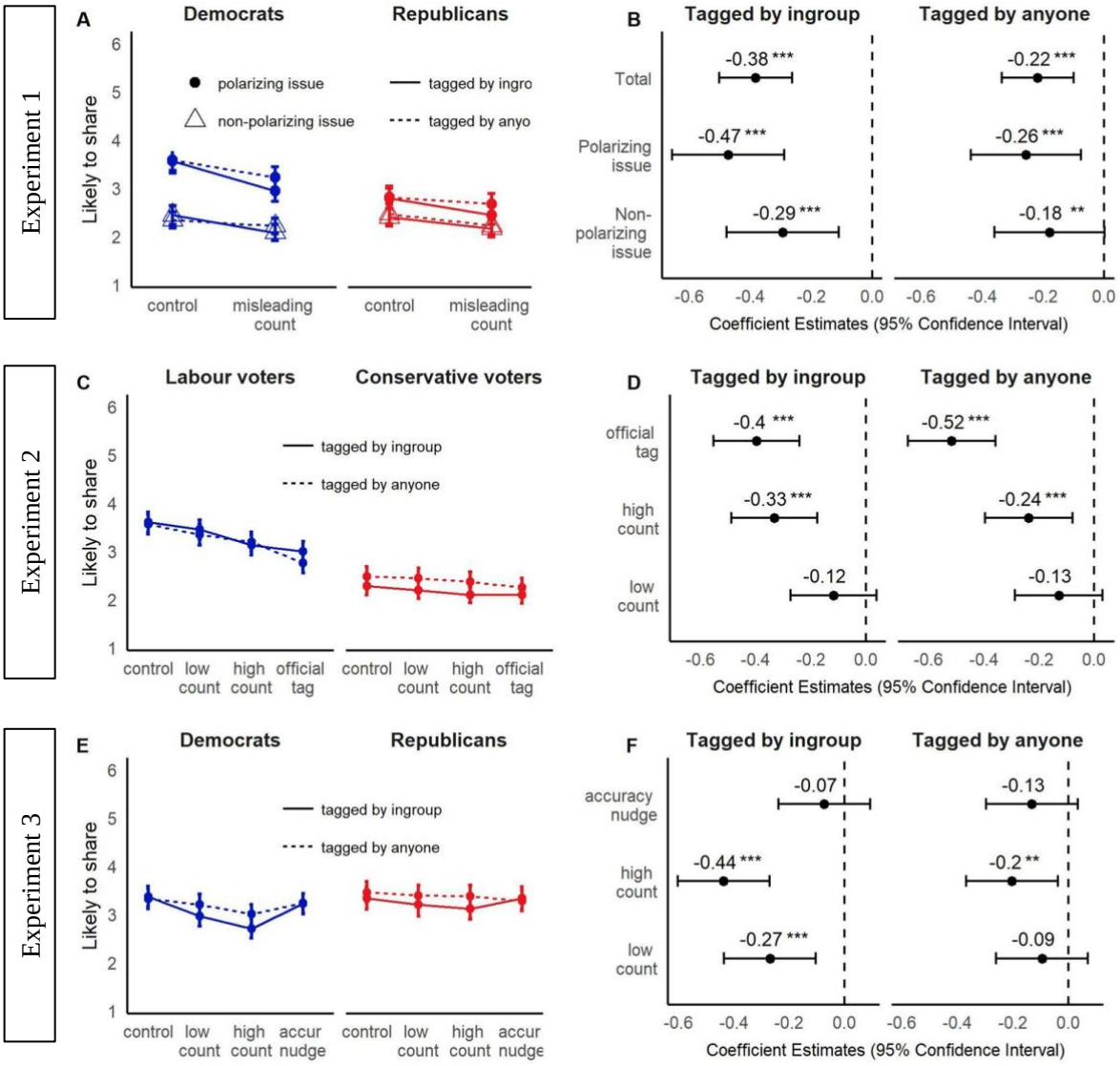

Figure 2. Effect of the intervention across experiments. (A, C, E) Likelihood of sharing social media posts on polarizing issues (Exp. 1-3) and non-polarizing issues (Exp. 1) in response to the *Misleading* count (Exp.1-3), the official Twitter tag (Exp. 2) and the accuracy nudge (Exp. 3) compared to control by group and condition ("tagged by in-group" and "tagged by anyone"). In Exp. 1, the *Misleading* count was always 30% of the *Like* count. In Exp. 2 and 3, the *Misleading* count was presented in two conditions: high count (80% of the *Like* count) and low count (20% of the *Like* count). The absolute *Misleading* count was two orders of magnitude higher in Exp. 1 as compared to Exp. 2 and 3 (e.g., 10k *versus* 100). Error bars represent 95% confidence intervals. **(B, D, F)** Coefficient estimates of the contrast between each intervention compared to control for social media posts relevant to polarizing (Exp. 1-3) and non-polarizing issues (Exp. 1) by condition ("tagged by in-group" and "tagged by anyone"). Error bars represent 95% confidence intervals.

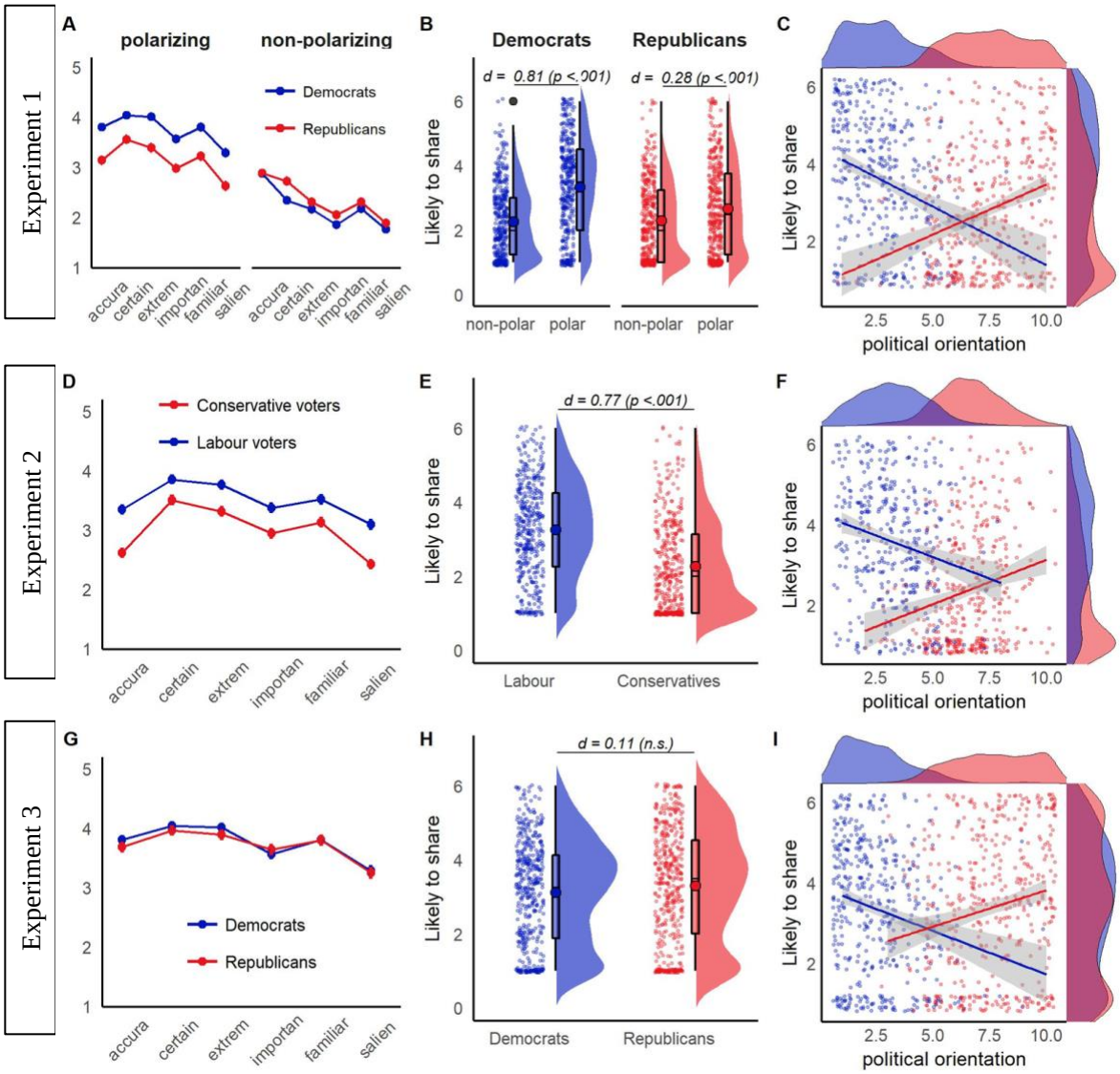

Figure 3. Perceived accuracy, attitude strength, and likelihood of sharing social media posts across experiments. (A, D, G) Perceived accuracy, attitude strength (certainty, extremity, and importance), familiarity, and salience of the used social media posts by group (Exp. 1-3) and type of issue (polarizing and non-polarizing) (Exp. 1) as tested in pilot studies (Exp 1: N = 370; Exp 2: N = 80; Exp 3: N = 234). **(B, E, G)** Likelihood of sharing social media posts on polarizing issues (Exp. 1-3) and non-polarizing issues (Exp. 1) as a function of political affiliation. **(C, F, I)** Extreme political orientation was associated with an increased likelihood of sharing social media posts in the control condition (no interventions) across samples and political groups (liberals in blue and conservatives in red). Notably, U.S. samples in Exp. 1 and 3 were more polarized in terms of political orientation compared to the UK sample in Exp. 2.