This working paper has not yet been subject to formal peer review.

Version 1: August 23, 2022

Version 2: December 12, 2022

Promoting engagement with social fact-checks online

Mohsen Mosleh^{*1,2}, Cameron Martel^{*2}, Dean Eckles^{2,3}, David G. Rand^{2,3,4}

Corresponding author: m.mosleh@exeter.ac.uk

^{*}Authors contributed equally

¹SITE (Science, Innovation, Technology, and Entrepreneurship) Department, University of Exeter Business School, Exeter EX4 4PU, UK; ²Sloan School of Management, Massachusetts Institute of Technology, 02142 Cambridge, MA, USA; ³Institute for Data, Systems, and Society, Massachusetts Institute of Technology, 02142 Cambridge, MA, USA; ⁴Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 02139 Cambridge, MA, USA

Social media users who post misinformation rarely engage with corrections from other users. We investigate the causal effect of the social relationship between the corrector and corrected user on engagement with such corrections. In a Twitter field experiment and an Amazon Mechanical Turk conceptual replication, we used human-looking bots to correct users who shared false news. We varied whether the corrector bot identified as a Democrat or Republican; and whether the corrector bot followed the user and liked three of their tweets the day before issuing the correction (creating a minimal social connection). Surprisingly, we did not find evidence that shared partisanship increased engagement with the correction (although overall, conservatives were more likely to engage negatively, versus positively, with corrections). Social connection, conversely, increased engagement with corrections from co-partisans, while effects for counter-partisans were ambiguous. A final survey experiment demonstrated a generalized norm of responding, where people feel more obligated to respond to people who follow them -- even outside the context of misinformation correction.

Introduction

There is widespread concern about the spread of false claims on social media – regarding, for example, candidates in the 2016 U.S. Presidential Election¹, the COVID-19 pandemic², and fraud in the 2020 U.S. Presidential election³. As a result, a great deal of research in recent years has investigated ways to combat the spread of misinformation on social media platforms⁴⁻⁶.

One such mechanism that has received considerable attention is *social correction*, wherein social media users correct one another. Numerous survey experiments indicate that corrections in general reduce belief in the debunked claims (e.g., ref 7), and almost never backfire by increasing belief in the debunked claim⁸ (although corrections can have negative impacts on subsequent sharing⁹). Furthermore, an added benefit of *social* correction in particular is that social corrections are observed not only by the person being corrected, but also by third parties (e.g., the followers of the person who is being corrected) — and survey experiments indicate that social corrections often cause these third party observers to update their beliefs (e.g., ref 10). Social corrections also have the advantage of centering agency in the hands of social media users, rather than making the platforms the arbiters of truth.

Here, we use field and survey experiments to shed new light on social correction by examining what factors cause users to ignore versus engage with social corrections that they receive. Engagement and reciprocal communication have been proposed as necessary aspects of deliberative discourse^{11,12}, and user-to-user interaction is a prerequisite for any attempt at collaborative discussion¹³. Furthermore, whereas typical one-off approaches for correcting misinformation (e.g., fact-checks, general warnings) have reliable but small effects¹⁴, rebuttal mechanisms utilizing increased engagement and argumentation may be more effective at promoting sustained learning and belief updating (e.g., as they are in science¹⁵; and in politics¹⁶). However, reply rates to social corrections on Twitter have observationally been found to be relatively low – one study, for example, found that 73% of social corrections were ignored¹⁷. Assessing what types of social correction are likely to elicit a reply and initiate dialogue, rather than simply be ignored, is integral to promoting productive and helpful social debunking online. Thus, in the current work, we examine how to promote the crucial first step of initiating preliminary engagement.

While engagement versus non-engagement (‘ignoring’) is directly observable from on-platform behavior, it is more difficult to ascertain the resulting beliefs of individuals who engage

with corrections. Nonetheless, we additionally attempt to investigate whether individuals who engage with corrections in our experiment update their beliefs or resist contradictory information (see ref 18). Inferring and assessing the valence of initial correction engagement may elucidate how promoting replies to fact-checks could translate to more substantive downstream corrective or persuasive approaches, once the hurdle of users ignoring fact-checks has been cleared.

Our experiments focus on two factors which prior observational work suggest may be important for eliciting engagement with corrections: partisan alignment between the corrector and the corrected user, and social connection between the corrector and the corrected user. A large body of literature shows that people tend to dislike and distrust counter-partisans compared to co-partisans (for reviews, see refs 19, 20), and preferentially follow co-partisans on social media²¹. Accordingly, evidence suggests that partisanship may also impact correction efficacy. One survey experiment found that corrections from Republicans were more effective at correcting climate misinformation believed by fellow Republicans²². Other research suggests that debunking messages from unlikely partisan sources may be more impactful upon both co-partisans and counter-partisans²³ or that partisan source may have little effect on belief updating²⁴. Given that partisan Twitter users tend to selectively share and write fact-checking messages which support their own party and rebut counter-partisans^{25,26}, it is important to experimentally investigate whether shared partisanship is a pre-requisite for engagement with social corrections. If so, this would impose a substantial limitation on the reach of such corrections.

Separately, observational research suggests that Twitter users are more likely to respond to corrections¹⁷, and to have responses which are positive and indicative of correction acceptance²⁷, when the correction comes from a user with which they have social connections (e.g., both corrector and corrected follow each other). However, since this prior work is observational, it remains unclear whether this effect is causal. Furthermore, even if it is causal, it is unclear whether it is necessary for there to be an actual pre-existing offline relationship, or if the kind of minimal virtual social connection that social media can facilitate is sufficient to increase engagement with corrections.

Here, we experimentally test the impact of partisan alignment and virtual social connection on engagement with corrections. We conduct a randomized experiment on Twitter in which we use human-looking bot accounts to deliver corrections to users that shared false claims. We vary whether the corrector identified as Democrat or Republican; and whether the corrector followed

the user and liked three of their tweets the day before issuing the correction (thus creating a minimal social relationship). We also examine whether partisanship or partisan extremity predicts engagement with correction messages or moderates the effects of our treatment conditions. We then replicate our results, and shed some light on mechanism, in a follow-up survey experiment on Amazon Mechanical Turk. Finally, we conduct a third experiment examining the role of communicative reciprocity norms on the relationship between minimal social connections and responding to corrections of misinformation.

Results

Twitter field experiment

To study causal effects of the corrector–correctee relationship on engagement with social corrections in the field, we first identified Twitter users who had written posts containing links to one of 11 articles which were rated false by the fact-checking website Snopes.com (see Supplementary Table 1 for list of the 11 articles). We then publicly replied to the false tweet with a corrective message that questioned the veracity of the claim and provided a link to the fact-checking article debunking the claim (see Figure 1). Although research has found no significant effect of the tone of corrective messages on intentions to reply versus ignore^{28,29}, out of caution we used non-confrontational language in our corrections (See Supplementary Table 2 for variations of the corrective messages we used).

Figure 1. Correction format. (a) Sample false tweet (text blurred out to protect privacy), (b) corrective message, and (c) user reply from the Twitter field experiment.

The corrective replies were made using bot accounts designed to look human, and the key manipulation in our experiment involved varying the corrector–correctee relationship in a two-by-two experimental design. First, we varied the apparent partisanship of the bot account (see Figure 2a,b), thereby experimentally varying whether the user and the correcting bot were co-partisans or counter-partisans. Second, we varied the presence versus absence of social connection between the user and correcting bot (Figure 2c,d). In the social condition, the correcting bot followed the user and liked their last three tweets on the day before the correction was delivered; in the baseline condition, conversely, the user and the bot had no prior interaction.

Figure 2. Experimental conditions. Design of partisan, social bot accounts. (a) Democrat bot account. (b) Republican bot account. (c) Social condition in which bot followed user and liked 3 of their tweets. (d) Baseline condition in which bot had no prior contact with user. Because candidate names were mentioned directly in the bots' usernames, even users who did not click into a bot's profile would be able to clearly identify the bot's partisanship.

Differential message delivery failure rates across conditions

We targeted 2,000 Twitter users who had shared false content, randomized to be corrected by our bots over the course of 18 days. However, Twitter restricted our accounts after 14 days of issuing corrections, forcing us to discontinue the experiment earlier than planned and resulting in 414 users who we did not attempt to correct. In total, we sent 1,586 corrective messages, of which 1,456 were successfully delivered to the users (see Supplementary Table 3 for counts of engagement and positive and negative responses across conditions). The remaining 130 corrective messages were sent but unable to be delivered, due to users blocking our bot accounts, users' accounts being protected, or the targeted tweets having been deleted.

Importantly, the fraction of participants to whom corrections were not successfully sent varied across experimental conditions. In particular, there was significant variation across conditions in the probability of receiving an error indicating that the tweet to which we were trying to reply was “deleted or not visible to you” (baseline co-partisan: 3.79%, baseline counter-partisan: 4.50%, social co-partisan: 6.43%, social counter-partisan: 17.96%; Fisher's exact test, $p < 0.001$). The message failure rate was specifically substantially higher in the social counter-partisan condition, likely driven in large part by users blocking the counter-partisan bot after it followed them. When testing the difference in message failure rates across the other three conditions, the difference was not significant (Fisher's exact test, $p = 0.200$).

This differential message failure rate creates an unforeseen challenge for making causal inferences about the differences in engagement involving the social counter-partisan condition. That is, differences between engagement in the social counter-partisan condition and the other conditions could be driven by selection effects, rather than actual causal effects. No such challenge exists, however, when making comparisons between the other three conditions (as message delivery failure rates did not differ).

Analyses unaffected by differential message delivery failure

Given the causal inference issues associated with the social counter-partisan condition, we deviate from our pre-registered analysis plan and begin by considering a model that only includes the three conditions for which the error rate was not significantly unbalanced (i.e., omitting $N = 329$ in the social counter-partisan condition, leaving $N = 1,127$).

First, focusing on the co-partisan condition, we find a significant positive effect of social connection on engagement ($b = 0.066$, $p = 0.042$), such that being followed by a co-partisan and having them like three previous tweets prior to delivering the correction increased the user's likelihood of engaging with that correction.

We next investigated the effect of experimental conditions on both positive and negative initial engagement with our corrective messages (for descriptive statistics, see Supplementary Table 3). We recruited workers from Amazon Mechanical Turk to classify each reply tweet as either accepting or resisting the correction (see Methods and Supplementary Information Section 1.6). We defined positive engagement as a reply that expressed belief updating, liking or

retweeting the correction, or deleting of the tweet containing false news. We defined negative engagement as whether the reply resisted or challenged the correction.

We found that in the co-partisan condition, social connection increased negative engagement ($b=0.052$, $p=0.043$) but had no significant effect on positive engagement ($b=0.014$, $p=0.571$). However, the effect of prior social connection on negative versus positive engagement was not significantly different ($p=0.212$). Thus, although social connection significantly increased negative engagement and had a non-significant (but positive) effect on positive engagement in the co-partisan conditions, the lack of significance in the difference between these two effects means that we do not find evidence that the social condition increased negative engagement *more* than it increased positive engagement conditional on co-partisanship. This absence of significant difference between the effect of social connection on negative versus positive engagement should, however, be interpreted with some caution – we do indeed find evidence that social connection increases negative engagement, and do not precisely estimate whether there is no difference between the effect on negative versus positive engagement. Altogether, then, these results remain opaque as to whether prior social connection increases negative engagement in particular, or increases both negative and positive engagement, in the co-partisan conditions.

Second, focusing on the baseline condition, we find no significant effect of shared partisanship on engagement ($b=-0.003$, $p=0.916$; 95% CI on null $[-0.064, 0.057]$), such that users were no more likely to engage with corrections from co-partisans compared to counter-partisans. We also did not find a significant effect of shared partisanship in the baseline condition on either positive engagement ($b=0.000$, $p=0.990$; 95% CI on null $[-0.046, 0.045]$) or negative engagement ($b=-0.003$, $p=0.990$; 95% CI on null $[-0.049, 0.043]$).

Importantly, the social connection effect on engagement was significantly larger than the shared partisanship effect (Wald test, $\chi^2(1)=4.56$, $p=0.033$), providing some evidence that prior social connection has a larger effect on promoting correction engagement than shared partisanship.

Addressing differential message delivery failure in the social counter-partisan condition

We did not anticipate unsuccessful correction delivery *a priori*, and thus our pre-registration (https://aspredicted.org/blind.php?x=F4C_3Y5) did not specify how we would address the missing data that arose from users not receiving a correction (given that our outcome is whether or not the user responded to the correction, when no correction was sent, this outcome

is undefined resulting in missing data). In lieu of having a pre-registered plan for addressing missing data, we use two different approaches that allow us to ask different questions about the effect of social connection. First, we ask the practical question of which correction strategy leads to the highest overall rate of engagement with the correction on Twitter (regardless of successful correction delivery). To answer this question, we consider all users to whom we sent corrections, regardless of whether the correction was delivered successfully (and treat users whom we could not message as not engaging with the correction). Second, we ask the psychological question of how social connection affects whether users respond to correction messages, conditional on receiving them. To answer this question, we employed several analyses examining participants for whom the correction was successfully delivered. Given that none of this was specified one way or the other in our pre-registration, none of these analyses should be considered pre-registered.

Our reported analyses proceed as follows. First, to address the practical question of which correction approach would generate the highest level of engagement if deployed on Twitter, we filled in missing data with 0s (i.e., counted users who did not receive a correction as not engaging with the correction), and included all users regardless of successful correction message delivery. Second, we conducted analyses attempting to answer the psychological question of the effects of actually receiving a correction message. One such analysis simply excluded users to whom corrections were not successfully sent. However, the differential rate of missing data across conditions impairs our ability to make causal inferences using this analysis. Therefore, we also conducted an auxiliary analysis that sought to estimate the effect of all conditions among users who would have received a correction had they been assigned to the social counter-partisan condition. In particular, we took a principal stratification approach^{30,31} to infer the probability that data from each user in the three conditions with low missing data rates would have been not missing had the user been assigned to the condition with the high rate of missing data, and ran a model weighting users according to this probability.

Analysis including users for whom message delivery failed

First, we conduct an analysis that allows us to answer the practical question of the overall intervention efficacy of our correction-delivering bots. We used linear regression predicting whether the user engaged (0: user ignored or error message was received (likely due to blocking), 1: user replied to, retweeted, or liked the corrective message or deleted the tweet containing false

news) with the corrective message using the partisanship of the user (Democrat versus Republican; z -scored), partisan agreement between the user and corrector account (i.e., if user and corrector accounts were co-partisan or counter-partisan; z -scored; this deviates slightly from our pre-registration in which we planned to include bot partisanship rather than partisan agreement – see Supplemental Information Section 1.2), prior social connection (-0.5: baseline condition, 0.5: social condition), and all interactions (total $N=1,586$). For robustness and as pre-registered, we calculated p -values both using the standard asymptotic approach, and using Fisherian randomization inference³². With this specification, we did not find evidence of a significant effect of social condition ($b=0.027, p=0.208$), shared partisanship ($b=0.016, p=0.125$), or their interaction ($b=0.028, p=0.185$) on probability of engagement (see Figure 3a,d). We also did not find an association between user partisanship and probability of engagement ($b=0.005, p=0.612$). Further, we did not find an effect of experimental conditions on positive engagement (social condition $b=0.002, p=0.889$; shared partisanship: $b=0.004, p=0.586$; interaction $b=0.008, p=0.629$) or negative engagement (social condition $b=0.025, p=0.138$; shared partisanship: $b=0.012, p=0.147$; interaction $b=0.021, p=0.214$). We do find evidence of a partisan asymmetry in engagement valence. More conservative users were more likely to engage negatively with the correction ($b=0.019, p=0.025$) and nominally less likely to engage positively with the correction ($b=-0.013, p=0.093$), such that more conservative users were overall more likely to express negative than positive engagement ($p=0.005$).

Analysis excluding users for whom message delivery failed

Next, we conduct analyses focused on answering the psychological question of how users respond to corrective messages conditional on receiving them. We first looked at the effect of experimental conditions on whether the user responded to versus ignored the correction, excluding users to whom corrections were not successfully sent ($N=1,456$; Figure 3b,e; for engagement descriptive statistics by condition, see Supplementary Table 3). Here, we found a significant positive effect of prior social connection on the likelihood of the user engaging with the correction ($b=0.052, SE=0.023, t(1455)=2.242, p_{reg}=0.025, p_{FRI}=0.017$). Surprisingly, we did not find a significant effect of shared partisanship on engaging with the correction ($b=0.009, SE=0.012, t(1455)=0.748, p_{reg}=0.454, p_{FRI}=0.435$; 95% confidence interval on null effect [-0.01, 0.03]). All other independent variables and their interactions were non-significant ($ps>0.1$), and in particular

the interaction between prior social connection and shared partisanship was not significant, $b=0.016$, $SE=0.023$, $t(1455)=0.690$, $p_{reg}=0.490$, $p_{FRI}=0.475$ (See Supplementary Table 8 and Supplementary Information Figure 2). Our results are also robust to using Bayesian multi-level models (See Supplementary Table 12).

For positive engagement, we do not find a significant effect of prior social connection ($b=0.013$, $SE=0.017$, $t(1455)=0.782$, $p_{reg}=0.435$, $p_{FRI}=0.430$), shared partisanship ($b=0.000$, $SE=0.009$, $t(1455)=0.006$, $p_{reg}=0.995$, $p_{FRI}=0.996$), or their interaction ($b=0.001$, $SE=0.017$, $t(1455)=0.039$, $p_{reg}=0.955$, $p_{FRI}=0.961$). For negative engagement, we also do not find a significant effect of shared partisanship ($b=0.009$, $SE=0.009$, $t(1455)=0.945$, $p_{reg}=0.345$, $p_{FRI}=0.303$), and do not find a significant interaction between shared partisanship and prior social connection ($b=0.015$, $SE=0.018$, $t(1455)=0.824$, $p_{reg}=0.410$, $p_{FRI}=0.410$). We do, however, find a significant main effect of social connection on negative engagement ($b=0.038$, $SE=0.018$, $t(1455)=2.111$, $p_{reg}=0.035$, $p_{FRI}=0.033$). Of note, the effect of prior social connection on negative versus positive engagement was not significantly different ($p_{reg}=0.325$, $p_{FRI}=0.395$). As in our analyses examining the effect of social connection amongst the co-partisan conditions only, we again do not find evidence that the social condition increased negative engagement *more* than it increased positive engagement (for full results, see Supplementary Tables 9-11, 13-14). However, we again emphasize that this absence of significant difference between negative and positive engagement should be interpreted with caution, as we still find a directionally stronger effect of social connection increasing negative engagement more so than positive engagement.

We also again find some evidence of a partisan asymmetry in engagement valence. User partisanship is not associated with overall engagement ($b=0.001$, $SE=0.012$, $t(1455)=0.064$, $p_{reg}=0.949$), and we do not observe any significant interactions between user partisanship and either of our experimental conditions ($ps>0.075$). However, we do find that more conservative users were marginally less likely to engage positively with the correction ($b=-0.017$, $SE=0.009$, $t(1455)=-1.975$, $p_{reg}=0.050$), and were also marginally more likely to engage negatively with the correction ($b=0.018$, $SE=0.009$, $t(1455)=1.944$, $p_{reg}=0.052$) - such that more conservative users were more likely to express negative than positive engagement ($p_{reg}=0.006$). These results provide preliminary evidence of a partisan asymmetry in responses to corrective information, although it is important to keep in mind that in terms of partisanship our sample is not representative of either Twitter users or Americans more generally.

Figure 3. Prior social connection increases engagement with corrections in our Twitter field experiment, particularly given shared partisanship. Probability of engagement with the corrective message in the field experiment is shown across experimental conditions. Panels (a) and (d) show, respectively, the probability of engagement and effect sizes from the linear model when including users for whom the correction message failed to send. Panels (b) and (e) show, respectively, the probability of engagement and effect sizes from the linear model when excluding users for whom the correction message failed to send. Panels (c) and (f) show, respectively, the probability of engagement and effect sizes from the linear model when weighting observations in conditions other than the Social Counter-partisan condition based on their predicted probability of successful delivery of the correction message had they been in the Social Counter-partisan condition (principal stratification). Error bars reflect 95% confidence intervals.

We also perform several additional exploratory analyses examining whether user partisanship *extremity* moderates the effects of prior social connection or shared partisanship on engagement. Is it the case that more politically extreme users were less receptive to corrections from counter-partisans, and does this change the effect of social connection?

We performed the same model as above, except replacing user partisanship with user partisanship extremity (i.e., the absolute value of z -scored user partisanship). We find a significant main effect of user partisan extremity on engagement, such that more politically extreme users were more likely to engage with the correction ($b=0.049$, $p<.001$); however, we do not find any significant interactions between user partisanship extremity, prior social connection, and/or shared partisanship ($ps>.357$; see TI Table 68).

We also investigated whether there could be potential non-linear effects of partisan extremity on engagement by experiment conditions. We classified users into four bins based on partisan extremity, and then compared the most extreme users (those in the highest extremity bin) to less extreme users (those in the three lower extremity bins); maximal extremity $N=335$; lower extremity $N=1,115$). In a linear regression predicting engagement from social condition, shared partisanship, and a dummy for partisan extremity (0=lowest 3 bins, 1=highest bin), we find a significant three-way interaction ($b=0.136$, $p=.013$; see Figure 4; SI Table 71). This three-way interaction is also significant when including all users in our analysis, including those who never received a correction ($b=0.064$, $p=.003$; See SI Table 72, Supplementary Figure 7).

Figure 4. Effect of prior social connection and shared partisanship on correction engagement, by user partisan extremity (binned: low extremity = lowest 3 bins, high extremity = highest bin). Error bars reflect 95% confidence intervals.

Decomposing this interaction, we find that for lower political extremity users, there is a significant effect of prior social connection on increasing engagement ($b=0.055$, $p=.033$), and no

main effect of, or interaction with, shared partisanship ($ps > .477$). For the most politically extreme users, however, the pattern was quite different: instead of significant main effects for either prior social connection or shared partisanship ($ps > .360$), there was a significant interaction between the two ($b = 0.118$, $p = .020$) such that prior social connection increased engagement for co-partisans, but decreased engagement for counter-partisans. This suggests that for non-extreme partisans, prior social connection increases engagement for both counter-partisan and co-partisan exchanges – but that for the most extreme partisans, prior social connection increases engagement amongst co-partisans while decreasing engagement for counter-partisan exchanges.

Principal stratification analysis

Finally, we attempted to address differential message delivery failure rates across conditions via a principal stratification approach^{30,31}, which allows us to estimate the effects of our experimental conditions among users for whom the correction was successfully delivered by leveraging available information about users for whom the correction was *not* successfully delivered. Specifically, we used the participants in the social counter-partisan condition to fit a random forest model predicting probability of successful correction delivery. We used users' characteristics including number of favourites, followers, friends, lists, statuses, tweets in the past two weeks, replies in the past two weeks, and total replies, as well as estimated ideology partisanship (Democrat vs Republican) and partisan extremity as the set of features to predict whether the correction was successfully delivered (model AUC=0.66). We then used the random forest model to predict the probability of successful correction delivery for all users in the other three conditions, and re-ran our analysis excluding users for whom the corrections failed ($N = 1,456$) while additionally weighting observations by each user's predicted probability of successful correction delivery had they been assigned to the Social Counter-partisan condition. This helps us to address the causal inference challenge arising from our inability to know which participants in the other three conditions would have blocked us had they been assigned to the social-discordant condition. The outcomes using this approach are qualitatively equivalent to our prior analysis results. For overall engagement rate, we found a significant positive effect of social condition ($b = 0.052$, $p = 0.025$), did not find a significant effect of shared partisanship ($b = 0.008$, $p = 0.467$), and did not find an interaction between conditions ($b = 0.016$, $p = 0.486$; see Figure 3c,f) on the probability of engagement with corrective message. For positive engagement, we did not find a

significant effect of social condition ($b=0.014$, $p=0.413$), nor of shared partisanship ($b=-0.001$, $p=0.915$), and did not find interaction between conditions ($b=0.002$, $p=0.902$). For negative engagement, we found a significant effect of social condition ($b=0.038$, $p=0.038$), did not find a significant effect of shared partisanship ($b=0.009$, $p=0.306$), and did not observe an interaction between conditions ($b=0.014$, $p=0.443$). We also did not find a significant difference in the effect of social connection on negative versus positive engagement ($p=0.385$); though again, this null result should be interpreted with caution, as it is plausible that social connection may still differentially increase negative engagement more so than positive engagement, as our null effects were not precisely estimated.

Evidence that users were attending to partisanship

One possible explanation for why we do not find evidence of an effect of shared partisanship is that users simply did not notice (or did not believe) the partisanship identification of the corrector bots. The preferential blocking of counter-partisan bots in the Social condition, discussed at length above, is one indication that users were indeed attending to bot partisanship.

Additionally, we conducted a post-hoc analysis focusing on users in the social condition (where the bot followed the user prior to issuing the correction) and asking whether the users preferentially reciprocated tie formation with co-partisan bots. Consistent with other work on tie formation on Twitter²¹, we found that users were *much* more likely to follow-back co-partisan bots compared to counter-partisan bots (follow-back probability for co-partisans, 0.316; follow-back probability for counter-partisans, 0.134; $p<.001$; see Figure 5 and SI Tables 16, 17 and Supplementary Figures 5, 6). For comparison, the follow-back rates of Forbes top charities are only 8% on average³³ – indicating that these follow-back rates, particularly for the co-partisans, are very high. The difference in follow-back rate between co- and counter-partisan bots indicates that many users *were* attending to the corrector bot's partisanship, and the lack of effect of shared partisanship on correction engagement cannot easily be attributed to mere unawareness or disbelief in the bots' professed partisanship. We also examine whether there is descriptive evidence that users suspected our human-appearing bots were in fact bots. Reassuringly, we find that out of the 1,456 users who received corrections, only 4 replies include the word “bot.” Overall, then, we conclude that there is little evidence to suggest systemic suspicion of our bots. Furthermore, bot suspicion would likely be constant across our experimental conditions, since our accounts only

varied by initial social contact and expressed partisanship; thus, non-differential bot suspicion across conditions would not undermine our causal inference analyses.

Figure 5. Shared partisanship substantially increased follow-back rate in the social conditions. Probability of follow-back in social condition based on shared partisanship between the user and corrector account. Error bars reflect 95% confidence intervals.

Survey experiment 1

To shed further light on the psychological effects of social connection and shared partisanship on engagement among those who receive a correction message, we conducted a survey experiment on Amazon Mechanical Turk that largely reproduced the conditions of the field experiment, while preventing users from not receiving a correction¹. The survey experiment therefore allows us to assess the replicability of the findings from the field experiment in a context that does not have the causal inference issues faced in the field experiment. Furthermore, the

¹ We also conducted a similar pre-registered follow-up survey experiment ($N=948$) prior to survey experiment 1 – see SI Section 5: Survey Experiment S1a. Here, we also largely replicated our main field experiment results predicting correction engagement. We found that prior social connection increased engagement ($b=.098$, $p=.002$). We did not find evidence that shared partisanship significantly increased engagement ($b=.050$, $p=.126$), nor evidence for an interaction between social connection and shared partisanship ($b=.025$, $p=.694$). Interestingly, here we did find that both prior social connection and shared partisanship increased replies indicating belief updating ($ps<.003$) and did not find evidence that either condition affected replies indicating resisting information ($ps>.641$).

survey experiment provides the opportunity to preliminarily examine several potential mechanisms underlying the positive effect of minimal social connection on correction engagement. We pre-registered this experiment (https://aspredicted.org/ZHT_SQW; deviations from pre-registration are discussed in Supplementary Information Sections 2.1, 2.4, and 2.5).

We recruited $N=815$ participants from Amazon Mechanical Turk who reported that they had a Twitter account; 3 participants failed to answer the main engagement dependent variable item, and 4 participants did not indicate their partisanship, resulting in incomplete assignment of shared partisanship – thus, our final sample included 808 participants. We then asked participants to imagine that they were on a social media platform such as Twitter. Participants in the social condition were next instructed to imagine that a fictional user, Dave P. Scott, had followed them on Twitter and liked 3 of their recent Tweets, and were shown a fictionalized Twitter notifications screen containing this information, as well as the partisanship of Dave P. Scott (Democrat, Republican) as indicated by Dave P. Scott’s Twitter name (Figure 6a). Participants in the no prior social connection condition did not receive these stimuli. Participants in both conditions were then told to imagine that several days after sharing a political news article on Twitter, they received a public reply on their post from @DaveP_Scott. All participants then received a corrective message stylized as a tweet, which read, “I’m not sure about this article – it could be false. I found a link on Snopes that says this headline is false. <https://www.snopes.com>.” (Figure 6b). Finally, we asked participants, “Would you reply to the above message?”, with response options “*Yes, I would write a reply*” and “*No, I would not write a reply*”. Given that there was no actual post that the correction was responding to, we did not ask participants to actually write a reply. Participants were also asked several questions about their attitudes towards the corrector and correction message in order to investigate several potential mechanisms behind reply likelihood, as described in more detail below.

(a)

(b)

Figure 6. Survey experiment 1 stimuli. (a) Participants in the social condition viewed fictionalized Twitter notification screens, with corrections either from a Democrat user (left) or Republican user (right). Other notifications are blurred to focus the subject attention on the notification that is used as the manipulation. (b) Stylized corrective messages shown to participants (Democrat condition left; Republican condition right).

As our main analysis, we used linear regression to predict participant reply (0=No, 1=Yes) given the partisanship of the participant (7-point Likert-scale, 1 = *Strongly Republican*, 7 = *Strongly Democrat*; z -scored; addition of participant partisanship reflects a deviation from pre-

registered analysis plan, see Supplemental Information Section 2.1), partisan agreement between the participant and corrector account (-0.5: counter-partisan, 0.5: co-partisan²), prior social connection (-0.5: baseline condition, 0.5: social condition), and all interactions (for descriptive statistics of engagement by condition, see Supplementary Table 18). As in our field experiment, we found a significant positive effect of prior social connection on the likelihood of participants indicating that they would reply to the correction ($b=0.083$, $SE=0.035$, $t(807)=2.373$, $p=.018$; see Figure 7; see Supplemental Information Section 2.3 for full main analyses). Again consistent with the field experiment, we did not find a significant effect of shared partisanship on probability of replying to the correction ($b=0.028$, $SE=0.017$, $t(807)=1.608$, $p=.108$). We also found a significant effect of user partisanship on reply likelihood, such that participants who were more pro-Republican were less likely to reply ($b=-0.048$, $SE=0.017$, $t(807)=-2.739$, $p=.006$). Importantly, the interaction between prior social connection and shared partisanship was not significant ($b=0.043$, $SE=0.035$, $t(807)=1.242$, $p=.214$).

Figure 7. Social connection, but not shared partisanship, increases engagement intentions in our survey experiment. Likelihood of replying to corrective message by condition. Error bars reflect 95% confidence intervals.

² For participants who answered 4 = *Neutral* on the continuous 1-7 partisanship measure, we used a binary choice presidential candidate preference item to classify partisanship for the purposes of identifying partisan agreement.

As in our field experiment, we also performed exploratory analyses examining the role of participant partisan extremity. We performed the same analysis as above, except substituting participant partisanship with participant partisan extremity (absolute value of z -scored partisanship). Replicating our field experiment results, we found that more politically extreme participants were more likely to engage with the correction ($b=0.048$, $p=.006$), and did not find evidence of interactions between participant partisan extremity and our treatment conditions ($ps>.393$; see SI Table 73). However, we also investigated whether treatment effects may have differed for the most politically extreme users. From our 7-point partisanship measure, we binned participants into either lower-extremity (2-6 on scale; $N=586$) or high-extremity participants (1 or 7 on scale; $N=222$).

The results were strikingly similar to those observed in the field experiment (see Figure 8; SI Table 76). A linear regression predicting engagement by social condition, shared partisanship, a dummy variable for user partisan extremity (0=lower 3 bins, 1=highest bin) and all interactions yields a significant three-way interaction ($b=0.214$, $p=.006$).

Figure 8. Effect of prior social connection and shared partisanship on correction engagement, by user partisan extremity. Error bars reflect 95% confidence intervals.

Decomposing this interaction, we find that for lower-extremity participants social connection increased engagement ($b=0.101$, $p=.014$) while there was no significant main effect of, or interaction with, shared partisanship ($ps>.268$). Conversely, for high-extremity participants

there were no significant main effects of prior social connection or shared partisanship ($ps > .201$), but instead a significant interaction between the two ($b = .201$, $p = .002$) such that prior social connection increased engagement for co-partisans, but decreased engagement for counter-partisans.

Overall, the main findings from our field experiment replicated in our follow-up Mechanical Turk survey experiment³ - most importantly, we again find that prior social connection increases engagement with correction messages, particularly for co-partisans. Since all participants in our survey experiment actually received a correction, these results and their replication of our field experiment findings help to address concerns related to differential attrition in our Twitter experiment.

We also performed linear regression analyses for several potential mechanism-related dependent variables (rudeness of not replying; obligation to reply; comfort having discussion with corrector; trustworthiness of corrector; intelligence of corrector; perceived appropriateness of correcting misinformation; positive/negative affect towards corrector; likelihood corrector is a bot), again entering social condition, participant partisanship, partisan agreement, and all interactions as predictors. We found preliminary evidence that participants in the social condition felt more obligated to reply to the corrective message than participants in the baseline condition, $b = 0.300$, $SE = 0.147$, $t(807) = 2.031$, $p = .043$. However, this effect of social condition on feeling obligated to reply was no longer significant after adjusting for multiple comparisons (calculated q -value = 0.294; simulated q -value = 0.225; see Supplementary Information Section 2.7). We did not observe any difference between the social and baseline condition for any of our other potential mechanism variables, $ps > .138$ (see Supplementary Information Section 2.3).

³ We also conducted similar analyses, except (i) using only binary presidential candidate preference as partisanship, and (ii) only using social condition and partisan agreement as predictors, and repeatedly found a positive effect of prior social connection on reply likelihood; we also did not find an effect of partisan agreement on reply likelihood. See Supplementary Information Sections 2.4-2.6.

Interestingly, we do observe an effect of shared partisanship on positive (versus negative) feelings towards the corrector ($p = .001$, calculated q-value = .011, simulated q-value = .009) despite the fact that we do not find an effect of shared partisanship on probability of responding. This further demonstrates the apparent disassociation between promoting engagement and fostering positive feelings towards the corrector – as minimal social connection appears to increase engagement but not positive sentiment, whereas shared partisanship may increase positive sentiment without increasing engagement. We also find some suggestive evidence of effects of shared partisanship on perceived corrector trustworthiness ($p = .015$, calculated q-value = .112, simulated q-value = .091) and intelligence ($p = .010$, calculated q-value = .080, simulated q-value = .065), though these latter effects are no longer significant after adjusting for multiple comparisons. Ultimately, these results suggest that social connection is perhaps promoting engagement via a mechanism more closely related to establishing a norm or obligation to respond, rather than enhancing positive feelings or affect towards the corrector; whereas shared partisanship may promote the latter, without necessarily increasing engagement.

Survey experiment 2

In addition to replicating the results of the field experiment, our first survey experiment provided some tentative evidence that minimal social relationships may increase users' sense of obligation to respond to corrective messages. This observation offers a potential mechanism behind the field and survey findings that minimal social relationships increase the likelihood of engaging with corrections. However, the effect of the minimal social relationship on feelings of obligation was only barely significant, and did not survive correction for multiple testing. Therefore, in our final survey experiment, we conduct a pre-registered

(https://aspredicted.org/G8Q_413; see Supplementary Information Section 3.1) conceptual replication and follow-up assessment of the effect of minimal social relationships on the sense of obligation to respond. We also ask whether minimal social relationships may foster a more generalized norm of response that goes beyond misinformation correction.

To do so, we recruited $N=1,606$ participants on Amazon Mechanical Turk, screening for participants who reported having Twitter accounts. Participants were told to imagine they were on Twitter. Participants were then randomly assigned to either the prior social connection or no prior social connection condition, as in our first survey experiment. Next, participants were randomly assigned to either the correction or non-correction condition. Participants in the correction condition were told they received a public reply on a previously written post, stating “I’m not sure about this article – it could be false. I found a link on Snopes that says this headline is false. <https://snopes.com>.” Participants in the non-correction condition were told they received a public reply on a previously written post, instead stating “I’ll check out this article – I’d like to know more. You might also find this story interesting: <https://t.co/news-updates/>...” Finally, participants were randomly assigned to answer one of two potential key dependent variable items. One item assessed personal obligation to respond, replicating the first survey experiment: “How obligated are you to respond to Dave P. Scott’s reply?” 1 = *Not at all obligated*, 7 = *Extremely obligated*. The other item examined perceptions of a social norm of responding: “How much do you agree or disagree with the following statement? Most Twitter users would expect me to respond to Dave P. Scott’s reply.” 1 = *Strongly Disagree*, 7 = *Strongly Agree*.

As pre-registered, we used linear regression to predict personal obligation to reply ($N = 812$) by social condition (-0.5: baseline, 0.5: prior social connection), correction condition (-0.5: non-correction, 0.5: correction), and their interaction (for descriptive statistics by condition, see

Supplementary Table 59). We found that prior social connection increased personal obligation to reply ($b=0.592$, $SE=0.149$, $t(811)=3.985$, $p<.001$; See Figure 9a). We did not find an effect of correction condition on personal obligation to reply ($b=0.049$, $SE=0.149$, $t(811)=0.332$, $p=.740$), nor did we observe an interaction between social condition and correction condition on feelings of personal obligation to reply ($b=-0.154$, $SE=0.297$, $t(811)=-0.518$, $p=.605$)

Figure 9. Social connection increases reported personal obligation to reply and perceptions of a norm of reciprocal responding. (a) Reported personal obligation to respond to messages by condition. (b) Reported perceptions of reciprocal responding norm by condition. Bars indicate means, and raw data are shown with dots. Error bars indicate 95% confidence intervals.

We also used linear regression to predict perceptions of a norm of reciprocated responding ($N = 794$; for descriptive statistics, see Supplementary Table 60) by social condition, correction condition, and their interaction. Similar to our findings for personal obligation to respond, we found that prior social connection increased perceptions of a norm of responding ($b=0.245$, $SE=0.121$, $t(793)=2.019$, $p=.044$; See Figure 9b). Additionally, we also found that receiving a corrective reply increased perceptions of a norm of responding ($b=0.341$, $SE=0.121$, $t(793)=2.815$, $p=.005$). However, we did not observe an interaction between social and correction conditions on perceived norms of responding ($b=-0.056$, $SE=0.242$, $t(793)=-0.230$, $p=.818$).

Secondary analyses included partisanship (1-7 Likert; z -scored) as a potential moderator. Interestingly, we find that more pro-Republican participants were less likely to report a personal obligation to respond ($b = -0.215, p < .001$) and less likely to report a perceived norm of responding ($b = -0.107, p = .001$) overall - but we did not observe any interactions between partisanship and experimental conditions for either outcome variable. Thus, while more Republican participants may be less likely to feel obliged to respond and less likely to perceive a norm of responding, we do not find evidence that this partisan asymmetry exacerbates effects of prior social contact nor whether the reply was a correction. For full analyses, see Supplementary Information Section 3.3.

Altogether, these results support the idea that minimal social relationships increase feelings of obligation to respond, and also increase perceptions of a social norm of responding. Further, these results suggest that the effects of minimal social relationships on a communicative reciprocity norm are not unique to the domain of replying to corrections. Thus, there may be a more general social norm stipulating the need to respond to individuals based on prior social connections. Together with our field experiment and survey experiment 1, our findings suggest that minimal social connections may promote greater engagement with correction messages (particularly among co-partisans) because such relationships foster feelings of obligation to respond and a social norm of responding to those with whom a minimal connection is established.

Discussion

How does the relationship between corrector and correctee affect engagement with social corrections from other users on social media? Here, we used randomized experiments to shed light on this question.

Social connection

In both a Twitter field experiment and a survey experiment, we found evidence that “minimal” social connection online can increase the probability of engagement with a correction message from a co-partisan, and a follow-up survey experiment showed that minimal social connections online increase feelings of obligation to respond to public replies, and also increase perceptions of a general communicative reciprocity norm – even in outside the context of misinformation correction.

The effect of social connection when the correction message came from a counter-partisan was more complicated. In our Twitter field experiment, users in this condition were more likely to block our bots before the correction message was delivered – impairing our ability to assess whether they would have engaged with the correction had it been delivered. When including all users in our Twitter field experiment, we did not find evidence of prior social connection affecting correction engagement. When assessing treatment effects on only users who actually received correction messages, and in our conceptual replication survey experiment where all participants received a correction message, the results suggest that prior social connection may increase engagement with correction messages from counter-partisans for users who are not so politically extreme users – but for the most politically extreme users, social corrections may actually decrease engagement with corrections from counter-partisans.

When separating engagement in our Twitter field experiment into positive versus negative engagement, we did not find evidence of a significant difference in the effects of prior social connection on positive versus negative engagement, although our level of statistical power to detect such an interaction may have been limited. Thus, it may be the case that social connection increases negative engagement in particular, or increases engagement overall regardless of valence. Future work should investigate the valence of responses in more detail, as well investigating what happens over time when there is more sustained discussion – to what extent does initial negative engagement ultimately lead to more belief updating than simply ignoring the correction (due to stimulating more discussion with the corrector and/or other observers)?

Our results suggest that social corrections on social media, when used appropriately, may help to combat misinformation. Social media allows for the easy creation of minimal virtual social connections. As our results demonstrate, these minimal virtual social connections may help facilitate increased engagement with corrective messages. Therefore, users aiming to correct

misinformation may benefit from first socially connecting with those they wish to correct – particularly when corrector and correctee are co-partisans, or when partisanship is not indicated by the corrector (e.g., as in survey experiment 2). Conversely, prior social connection may not be helpful – or may even be actively harmful – when corrections come from counter-partisans. Users may be more likely to block counter-partisan users as soon as they are followed by them – such that prior social contact from counter-partisans may not be effective in promoting overall engagement with corrections. And politically extreme users who do receive corrections may be more likely to ignore corrections from counter-partisans after social connection. It is important to note, however, that these perverse effects may be exacerbated in our experimental setting relative to typical engagement on Twitter because our bots so strongly signaled their partisan commitments (e.g., referring to presidential candidates directly in their screen name). It seems plausible that social connection may effectively elicit greater overall engagement from counter-partisans when the corrector’s partisanship is less strongly signaled (or when the corrector is politically neutral). Future work should investigate this possibility.

Our results on social connection also provide potential guidance for effective fact-check delivery on social media platforms – for instance, platform-directed correcting bots or accounts may also benefit from only correcting misinformation from users they have already recently followed or interacted with online; particularly if such bots express co-partisanship with users, or perhaps do not display partisan affiliation at all. Such prior interactions may then increase engagement with corrections of misinformation, which in turn may help quell the spread of false news on social media.

From a theoretical perspective, our second survey experiment also sheds light on a general communicative reciprocity norm that goes beyond just corrective messages. Given that we find individuals in the social condition report a greater obligation to respond, and report that others would expect a response, it may be that there exists a general norm to reciprocate social contact on social media. Such a norm has been described, for example, in adolescent texting behavior³⁵. Our findings are also consistent with research on Twitter reply networks which find that social relationships, rather than discussion topics, generally condition reply likelihood³⁶. Given that we found these effects for both responding to corrective and non-corrective replies, our current findings reflect how a general communicative norm on Twitter may be utilized for more effective and engaging fact-checking and quelling of misinformation.

Partisanship

In addition to investigating the effect of social connection, we also investigated the causal effect of shared partisanship on correction engagement. Contrary to our expectations, we did not find evidence of a meaningful effect of shared partisanship on correction engagement in the baseline condition. We also found (subject to the caveats about differential message delivery failure discussed above) no evidence that shared partisanship moderated the effect of social connection, except for the most extreme partisans (for whom counter-partisanship decreased engagement in the social condition). Thus, for most users, shared partisanship appears to have surprisingly little effect on engagement with corrections.

That is not to say, of course, that shared partisanship had no effects – in addition to the interaction with social connection for extreme users, we also found across all users that there were large effects of shared partisanship on rates of follow-back and blocking. We also note that we do not precisely estimate a null effect of shared partisanship on correction engagement – rather, we do not find evidence of a shared partisanship effect, or interaction between shared partisanship and prior social connection (except for extreme users), in our field experiment or replication survey experiment.

We also found large differences in correction engagement based on partisanship itself. Right-leaning users were more likely to respond negatively (rather than positively) to correction messages than left-leaning users. In our survey experiments, right-leaning users were also less likely to engage with corrections at all, and less likely to report either a personal obligation to respond or a perceived norm of responding to replies. This suggests that there may be partisan asymmetries in receptivity to corrections, which may help to explain in part why numerous studies have found that conservatives are more likely to share misinformation than liberals³⁷⁻³⁹.

Limitations and future directions

The current research has several limitations. First, all bot accounts in the field experiment, as well as the fictional corrector stimuli used in the survey experiment, were portrayed as White male accounts. Prior research has demonstrated differential responses to White- and Black-

presenting accounts on Twitter (e.g., ref 40, as well as asymmetries in engagement with Black-versus non-Black-presenting Twitter troll accounts⁴¹). Future research should investigate the generalizability of our findings for non-White and non-male correctors, and may explore different demographic corrector-correctee dynamics in social correction interactions.

Second, our Twitter field experiment only targets users who shared one of 11 articles rated false by Snopes.com. These articles were selected using the following criteria: (i) focusing on a political and partisan subject, and (ii) having a Snopes.com fact-check article written in the 2 months prior to our field experiment. Future work should investigate how our results generalize to other false articles - for instance, false news that has yet to be fact-checked by professional fact-checkers; non-political false news; or false news from time periods other than that immediately preceding our experiment.

Third, Mechanical Turk samples are convenience samples that make no effort to be nationally representative, and the set of Twitter users who post links to debunked articles are also highly non-representative of the American public. Thus, we hesitate to make strong claims about the association between user partisanship and correction engagement, and instead focus on the effects of our experimental conditions of prior social connection and *shared* partisanship. Future work may examine whether partisan asymmetries exist in different populations of interest.

Additionally, our results are also limited in scope to examining Twitter – different engagement norms may exist on different platforms with distinct social tie formation mechanisms; for instance, ‘following’ on Twitter only necessitates a one-way social tie, whereas ‘friending’ on Facebook establishes a two-way tie. Likewise, blocking propensity and norms may differ across platforms, which may have implications for the efficacy of different bot designs (e.g., co-partisan versus counter-partisan social bots). Therefore, the extent to which prior social contact facilitates correction engagement, or engagement more generally, across platforms is an important direction for future research.

Fourth, although there is some observational evidence that corrections on social media which include a link to Snopes increase the likelihood that reshared misinformation would be deleted, such rumour cascades have been found to propagate even after accumulating hundreds of Snopes corrections^{42,43}. Therefore, it is almost certainly the case that although social ‘snoping’ messages can increase engagement with a correction message, misinformation can still spread and

propagate online despite such corrective engagement. Successfully targeting misinformation requires the combination of many different approaches.

Fifth, one of the major benefits of online social corrections is providing useful corrective information for third-party viewers¹⁰ – and increasing original poster engagement with corrections may further increase visibility and attention towards corrections. In our studies, we focused on engagement by the original tweeter rather than third-parties. Future work should investigate the second order effects of social connections and increasing original tweeter engagement with fact-checks on third-party viewers – indeed, it may be the case that increased engagement with corrective messages can increase overall third-party viewership of the correction.

Sixth, a follow-up analysis of the data reported in our paper suggests that social corrections may have unfortunate downstream consequences such as increasing the subsequent retweeting of low quality, toxic, or slanted content by the users who get corrected⁹. While such results may provide cause for concern regarding social corrections, such negative downstream effects may (i) be specific to low-attention sharing and may not extend to belief updating, and (ii) may potentially be mitigated by continued conversation and discourse following the initial correction message. Indeed, such additional conversation would be the next step in correction approaches relying on discussion and argumentation. Thus, future research should examine how different types of social corrections and conversational approaches may best advance both engagement with, and efficacy of, corrections. This is particularly important given our ambiguous results regarding whether the increased engagement fostered by prior social connection leads to initial belief updating or resisting of counter-attitudinal information. Finally, even if social corrections do either reduce the quality of subsequent sharing by the corrected user or prompt engagement which expresses resisting of corrective information, they may still be overall beneficial via correcting the beliefs of third-party observers¹⁰.

Conclusion

In sum, we have shown that minimal pre-emptive social contact makes Twitter users who post false news more likely to engage with corrections they receive, particularly from co-partisans. It is easier to ignore corrective information from total strangers.

Methods

Field Experiment

Participants. We selected 11 political articles that were rated as false by the fact-checking website Snopes.com in the two months prior to launching our Twitter field experiment (11-November-2019 to 3-January-2020; see Supplementary Information Table 1 for the full list of stories) and varied in their appeal to liberals and conservatives. We then used the source URLs of those articles to find users who posted links to those articles on Twitter (we focused on original tweets, rather than retweets). For each user who had shared a link to one of the false articles, we collected general account information (e.g., number of followers, number of followed accounts, number of tweets, and activities in the prior two weeks) and estimated political ideology using the accounts they followed on a continuous scale of $[-2.5, 2.5]$ ⁴⁴. As pre-registered, we removed ‘high status’ users (users with more than 10,000 followers, since such users are likely to receive large volumes of engagement, and thus we were concerned they might not notice our correction), as well as users for whom we could not estimate ideology because they did not follow any political accounts. Out of this set of users, we selected 2,000 users for inclusion in our study, in a fashion that maximized article political balance (i.e., we selected all liberal users, and then a random subset of conservative users. This led to 24.6% of our subjects being estimated to be liberal and 75.4% to be conservative).

Procedure. To deliver corrections to these users, we created six identical Democrat bots (Fig2a) and six identical Republican bots (Fig2b). The only difference between bots of a given partisanship was that each bot was given a different American male name. To appear authentic, we created the bot accounts roughly 3 months before interacting with users. As per ref 40, to make the accounts look more authentic, we purchased each account 1,000 followers (the followers did not have any engagement with our accounts). Before interacting with users, each bot account followed a set of accounts (i.e., major political figures and mainstream news outlets used in ref 44) aligned with their political ideology and randomly retweeted roughly 50 recent tweets from the timelines of those accounts. Bot accounts continued retweeting posts from political figures and news outlet every day during the course of the experiment. In the social condition, one day prior to sending the correction, the bot followed the to-be-corrected user and favorited the user’s three most recent

tweets (Fig2c). In the baseline condition, there was no prior interaction between the bot and the to-be-corrected user (Fig 2d).

We randomly assigned users within each block to one of 4 conditions based on the bots' partisanship (matched with user or mismatched with user) $\times$ prior social interaction with user (social, baseline), in a 2x2 between-subject design.

To balance subjects across experimental conditions and to improve the precision of our causal inference analysis, we performed randomized assignment by blocking⁴⁵. Specifically, we created relatively homogeneous blocks of users based on (i) number of followers, (ii) number of tweets in the past two weeks, (iii) number of mentions in the past two weeks, (iv) number of replies in the past two weeks, and (v) political ideology. We selected these covariates as they were the most predictive of the potential outcome in our pilot study. We used quantile transform of the count data to form the blocks as the counts were highly right-skewed.

To identify the causal effect of experimental condition on response to the corrective message, while also minimizing our risk of being sanctioned by Twitter (and thus needing to only send a small number of messages each day), we randomly assigned each set of blocks to a treatment day, in order to have a balanced number of users across experimental conditions receive corrections each day.

As the corrective message, we replied to the tweet containing the link to the false story with a short message and provided a link to the fact-checking article (each message used one of 12 similar wordings, see Supplementary Information Table 2 for the full list of message texts). To reduce the chance of corrected users noticing several similar corrective messages in the “tweets and replies” section of the bot profile, each bot would randomly retweet from the elite accounts it followed after sending each corrective reply.

We targeted 2,000 Twitter users who had shared false content, randomized to be corrected over the course of 18 days. However, our accounts got restricted by Twitter on day 14 and thus we had to stop the experiment at that point. Prior to the restriction, we sent corrective messages to 1,586 users, of which 1,456 were successfully delivered (failed delivery was caused by user's account being protected, users blocking the bot account, or by the tweet with the link to the false story having been deleted).

Our experimental setup was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) protocol #1907910465. We received a waiver of informed

consent as this was essential for the experiment to have ecological validity (i.e., for participants to not know they were part of an experiment), and the experiment posed no more than minimal risk to all persons. Minimal risk is typically evaluated by comparison with the risk people would be exposed to otherwise; being misinformed about a Twitter account's operator frequently occurs in the absence of our intervention. Furthermore, we did not debrief users. This is because we were unable to send private messages to the users, because they did not follow our accounts. We believed that publicly informing users (and their followers) that they had been part of an experiment on false news sharing would be a violation of their privacy, and that this would be a greater harm than not debriefing them given the minimal deception - the corrective tweet content was entirely accurate, the only deception was not identifying the corrector as a bot (which was needed in order to study social corrections). Furthermore, it is our understanding that this experiment was compliant with applicable legal frameworks, despite potentially violating Twitter's terms of service^{46,47}.

We pre-registered our analyses and sample size at https://aspredicted.org/F4C_3Y5. As described thoroughly in the main text, we did not achieve our pre-registered target of $N=2,000$ corrections because of interference from Twitter security. We note other deviations from our pre-registration below – namely, that we observed differential treatment delivery failure across conditions, resulting in our pre-registration being critically under-specified. Therefore, while we report our pre-registration (and attempt to adhere to it as closely as possible given its under-specification), we emphasize that all analyses should best be considered exploratory.

Analysis. Our main outcome variables are (i) whether the user engaged with our corrective message (i.e., whether the user replied to, retweeted, or favourited our corrective message, or deleted the tweet containing false news), (ii) whether the user engaged positively with our corrective message (belief updating, liking/retweeting the correction, deleting the original false news tweet), and (iii) whether the user engaged negatively with our corrective message (resisting corrective information). We used linear regression to predict whether the user engaged (0: user did not engage, 1: user replied to, retweeted, or liked our corrective message, or deleted the tweet containing false news) with the corrective message, based on the partisanship of the user (Democrat, Republican; z -scored), the partisan agreement of the user and corrector accounts (co-partisan, counter-partisan; z -scored), prior social interaction (-0.5: no prior interaction, 0.5:

corrector followed the user and liked three of their most recent tweets one day prior to sending the correction), and all interactions. This model deviates slightly from our pre-registered model, which used a dummy for bot partisanship rather than a dummy for bot concordance; we fully report the pre-registered model, which gives equivalent results, in the SI. We conducted analyses unaffected by differential message delivery failure (as described in the main text and below). First, given co-partisanship, we estimate the effect of prior social connection on engagement, positive engagement, and negative engagement. Second, given baseline condition, we estimate the effect of shared partisanship on engagement, positive engagement, and negative engagement. These analyses exclude the social counter-partisan condition (leaving $N=1,127$). Next, we include all users regardless of condition and successful correction delivery, scoring the outcome for users with unsuccessful delivery as 0 (did not reply).

To determine whether responses were positive or negative, we ran a survey on Amazon Mechanical Turk ($N=66$) in which workers were asked to rate 50 reply tweets each as either exhibiting ‘belief updating’ or ‘resisting information’ (see SI Reply Tweet Rating for full survey and wording). Each reply tweet was rated 9.978 times on average ($SD = 1.057$; min=6, max=12). Inter-rater agreement was moderate to good (Kendall’s $W=0.241$; $ICC(1,k)=0.646$; common guidelines interpret intra-class correlations greater than 0.60 as good, see ref 48). We also counted instances where the user favourited or retweeted the correction or deleted tweet containing false news as positive engagement. Also, including favouriting the correction as a positive response was not included in the pre-registration, as it did not occur to us beforehand, but it is clearly an indication of positivity towards the correction and thus we include it.

In total, there were 201 instances of negative engagement, 177 instances of positive engagement (also see Supplementary Table 15), and 1078 instances of a lack of engagement. As reported in the main text, the treatment effects did not differ significantly for positive versus negative responding. The full models for positive and negative responses can be found in the Supplementary Tables 9-11.

As described in the main text, we did not anticipate unsuccessful correction delivery *a priori*, and thus did not include in our pre-registration how to address differential failure rate due to counter-partisan bot blocking. In addition to the analysis described above, we conducted two additional exploratory analyses in order to attempt to robustly account for how differential failure rate undermining the ability to make causal inferences about treatment effects for users who

actually received treatment. First, we conducted analyses identical to those mentioned above, except excluding all users with unsuccessful treatment delivery. Second, to estimate the effect of all conditions *among users for whom the correction was successfully delivered*, we took a post hoc principal stratification approach^{30,31}.

Survey experiment 1

Participants. Based on the results of a pilot study with a similar design (effect size $f^2 = .0125$), we performed a power calculation using G*Power⁴⁹ for linear multiple regression: fixed model, single regression coefficient, and aiming for 90% statistical power. This power analysis yielded a suggested total sample size of 843. We therefore aimed to recruit 850 participants via MTurk. As pre-registered (https://aspredicted.org/ZHT_SQW), we also filtered for Twitter users as an eligibility criterion. At the beginning of the survey, we asked participants “What type of social media account do you use (if any)? Please answer honestly.” Participants could choose from *Facebook, Twitter, Snapchat, Instagram, WhatsApp, and Other*. Participants who did not select ‘Twitter’ were excluded from participating in the rest of the survey. Participants who selected all six options were also excluded, on suspicion that such participants were intentionally trying to pass our screening question. After excluding participants who began our survey but did not satisfy our Twitter eligibility criteria, we recruited a total of 808 eligible participants (289 female, $M_{\text{age}} = 35.92$).

Procedure. Participants were asked to “Imagine that you are currently on a social media platform such as Twitter.” Next, participants were randomly assigned to either the *social* or *baseline* correction condition, as well as either the *Democrat* or *Republican* correction condition. Participants in the social condition were then told “Imagine that Dave P. Scott has followed you on Twitter, and has liked 3 of your recent Tweets, as seen below. Participants in the social condition were then shown a fictionalized Twitter notifications screen, with follow and like notifications from Dave P. Scott. Participants in the Democrat condition viewed a Dave P. Scott profile with ‘Anyone but Trump 2020’ in the Twitter name; participants in the Republican condition viewed a Dave P. Scott profile with ‘Trump 2020’ in the Twitter name. See Figure 6.

Participants in the social condition were then told, “Now imagine that several days after sharing a political news article on Twitter, you receive the following public reply on your post from @DaveP_Scott. Remember that the user @DaveP_Scott started following you on Twitter several days ago, and has favorited several of your Tweets.” Participants in the baseline condition, who did not receive the stimuli shown in Figure 6 were told, “Now imagine that several days after sharing a political news article on Twitter, you receive the following public reply on your post from @DaveP_Scott. Suppose that the user @DaveP_Scott does not follow you on Twitter, and that you have never interacted with each other before.”

All participants then received a corrective message stylized as a Tweet, which read, “I’m not sure about this article – it could be false. I found a link on Snopes that says this headline is false. <https://www.snopes.com>.” Participants in the Democrat condition again saw ‘Anyone but Trump 2020’ in the Twitter name of Dave P. Scott; participants in the Republican condition again saw ‘Trump 2020’ in the Twitter name of Dave P. Scott. See Figure 6.

All participants were then asked our key measure of engagement with the corrective message: “Would you reply to the above message?” 1 = *Yes, I would write a reply*. 0 = *No, I would not write a reply*.

We then asked a multitude of questions assessing potential mechanisms as to why participants may be more likely to respond in the social condition. In random order, participants were asked the following questions, all assessed via a Likert-scale (1-7): “How rude would it be to not respond to Dave P. Scott’s reply?”; “How obligated are you to respond to Dave P. Scott’s reply?”; “How comfortable would you feel having an online discussion with Dave P. Scott?”; “How trustworthy is Dave P. Scott?”; “How smart is Dave P. Scott?”; “Do people in your community think it’s appropriate to correct misinformation posted by others on Twitter?”. Participants were then asked sequentially, “How negative or positive do you feel towards Dave P. Scott?”, and “How likely do you think it is that Dave P. Scott is a bot?”. Finally, we concluded with standard demographic questions (age, race, partisanship, etc.). Partisanship was assessed primarily via the question: “Which US political party do you identify with more strongly?” (1=*Strongly Republican*, 4=*Neutral*, 7=*Strongly Democrat*; reverse-coded for analyses). Partisanship was secondarily assessed via the presidential candidate preference question: “If you absolutely had to choose between only Clinton and Trump, who would you prefer to be the next President of the United States (1=*Hillary Clinton*, 2=*Donald Trump*).

Survey experiment 2

Participants. We recruited 1,606 participants on Amazon Mechanical Turk ($M_{\text{age}} = 36.90$, 644 female). As pre-registered (https://aspredicted.org/G8Q_413), we only recruited participants who reported using a Twitter account, as well as participants who passed two trivial attention checks at the beginning of the survey.

Procedure. Participants were asked to “Imagine that you are currently on Twitter.” Next, participants were randomly assigned to either the *social* or *baseline* correction condition, as well as either the *correction* or *non-correction* condition. Participants in the social condition were then told “Imagine that Dave P. Scott has followed you on Twitter, and has liked 3 of your recent Tweets, as seen below. Participants in the social condition were then shown a fictionalized Twitter notifications screen, with follow and like notifications from Dave P. Scott. Unlike in the prior survey experiment, no partisanship identifier was included in Dave P. Scott’s Twitter name.

Participants in the social condition were then told, “Now imagine that several days after sharing an article on Twitter, you receive the following public reply on your post from @DaveP_Scott. Remember that the user @DaveP_Scott started following you on Twitter several days ago, and has favorited several of your Tweets.” Participants in the baseline condition, who did not receive the prior social interaction stimuli, were told, “Now imagine that several days after sharing an article on Twitter, you receive the following public reply on your post from @DaveP_Scott. Suppose that the user @DaveP_Scott does not follow you on Twitter, and that you have never interacted with each other before.”

All participants then received a reply message stylized as a Tweet. In the correction condition, this reply read: “I’m not sure about this article – it could be false. I found a link on Snopes that says this headline is false. <https://www.snopes.com>.” In the non-correction condition, this reply read: “I’ll check out this article – I’d like to know more. You might also find this story interesting: <https://t.co/news-updates/>...”

Participants were then randomly assigned to answer one of our two key obligation measures. Participants assigned to the personal obligation dependent variable were asked: “How obligated are you to respond to Dave P. Scott’s reply?” 1 = *Not at all obligated*, 7 = *Extremely*

obligated. Participants assigned to the normative obligation dependent variable were asked: “How much do you agree or disagree with the following statement? Most Twitter users would expect me to respond to Dave P. Scott’s reply.” 1 = *Strongly Disagree*, 7 = *Strongly Agree*.

Finally, we concluded with standard demographic questions (age, race, partisanship, etc.). Partisanship was assessed via the question: “Which US political party do you identify with more strongly?” 1 = *Strongly Republican*, 4 = *Neutral*, 7 = *Strongly Democrat*.

Data availability

Full experimental materials and code can be found here: https://osf.io/nvk4u/?view_only=2ba8efe34cd647e18f69857bd1f85bbd. For confidentiality reasons, the Twitter data are only available upon request.

Code availability

Full experimental materials and code can be found here: https://osf.io/nvk4u/?view_only=2ba8efe34cd647e18f69857bd1f85bbd. For confidentiality reasons, the Twitter data are only available upon request.

Acknowledgements

Authors thank Ben Tappin for insightful comments on Bayesian analysis and Antonio Arechar for assistance with survey experiment data collection. The authors gratefully acknowledge funding from Google through Google Research Scholar Award, gifts from Google to support other research, and funding via the National Science Foundation Graduate Research Fellowship under Grant No. 174530.

References

- 1 Lazer, D. M. *et al.* The science of fake news. *Science* **359**, 1094-1096 (2018).
- 2 Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3), 337-348.
- 3 Pennycook, G. & Rand, D. G. Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *The Harvard Kennedy School Misinformation Review* (2021).

- 4 Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest* **21**, 103-156 (2020).
- 5 Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends in Cognitive Sciences* (2021).
- 6 Wittenberg, C. & Berinsky, A.J. Misinformation and its correction. in *Social Media and Democracy: The State of the Field* (eds Persily, N. & Tucker, J. A.) (Cambridge University Press, 2020).
- 7 Ecker, U. K., O'Reilly, Z., Reid, J. S. & Chang, E. P. The effectiveness of short-format refutational fact-checks. *British Journal of Psychology* **111**, 36-54 (2020).
- 8 Wood, T. & Porter, E. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* **41**, 135-163 (2019).
- 9 Mosleh, M., Martel, C., Eckles, D. & Rand, D. G. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment. in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1-13 (2021).
- 10 Bode, L. & Vraga, E. K. See something, say something: Correction of global health misinformation on social media. *Health Communication* **33**, 1131-1140 (2018).
- 11 Habermas, J. The theory of communicative action, Volume 2: Lifeworld and system: A critique of functionalist reason. (1985).
- 12 Graham, T. & Witschge, T. In search of online deliberation: Towards a new method for examining the quality of online discussions. *Communications* **28**, 173-204 (2003).
- 13 Honey, C. & Herring, S. C. Beyond microblogging: Conversation and collaboration via Twitter. in *2009 42nd Hawaii International Conference on System Sciences*. 1-10 (Ieee).
- 14 Walter, N. & Murphy, S. T. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs* **85**, 423-441 (2018).
- 15 Osborne, J. Arguing to learn in science: The role of collaborative, critical discourse. *science* **328**, 463-466 (2010).
- 16 Jerit, J. Issue framing and engagement: Rhetorical strategy in public policy debates. *Political Behavior* **30**, 1-24 (2008).
- 17 Hannak, A., Margolin, D., Keegan, B. & Weber, I. Get back! You don't know me like that: The social mediation of fact checking interventions in twitter conversations. in *Proceedings of the International AAAI Conference on Web and Social Media*. (2014)
- 18 Prasad, M., Perrin, A. J., Bezila, K., Hoffman, S. G., Kindleberger, K., Manturuk, K., & Powers, A. S. (2009). "There must be a reason": Osama, Saddam, and inferred justification. *Sociological Inquiry*, 79(2), 142-162.
- 19 Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* **22**, 129-146 (2019).
- 20 Iyengar, S., Sood, G. & Lelkes, Y. Affect, not ideology: a social identity perspective on polarization. *Public Opinion Quarterly* **76**, 405-431 (2012).
- 21 Mosleh, M., Martel, C., Eckles, D. & Rand, D. G. Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences* **118** (2021).

- 22 Benegal, S. D. & Scruggs, L. A. Correcting misinformation about climate change: The impact of partisanship in an experimental setting. *Climatic Change* **148**, 61-80 (2018).
- 23 Berinsky, A. J. Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science* **47**, 241-262 (2017).
- 24 Swire, B., Berinsky, A. J., Lewandowsky, S. & Ecker, U. K. Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science* **4**, 160802 (2017).
- 25 Shin, J. & Thorson, K. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication* **67**, 233-255 (2017).
- 26 Allen, J., Martel, C. & Rand, D. G. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. in *CHI Conference on Human Factors in Computing Systems*. 1-19.
- 27 Margolin, D. B., Hannak, A. & Weber, I. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication* **35**, 196-219 (2018).
- 28 Martel, C., Mosleh, M. & Rand, D. G. You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online. *Media and Communication* (2021).
- 29 Bode, L., Vraga, E. K. & Tully, M. Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review* (2020).
- 30 Ding, P. & Lu, J. Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 757-777 (2017).
- 31 Frangakis, C. E. & Rubin, D. B. Principal stratification in causal inference. *Biometrics* **58**, 21-29 (2002).
- 32 Fisher, R. A. *The design of experiments*. (Macmillan, 1937).
- 33 Sun, S., & Rui, H. Link formation on twitter: the role of achieved status and value Homophily. in *Proceedings of the 50th Hawaii International Conference on System Sciences* (2017).
- 34 Rathje, S., Van Bavel, J. J. & van der Linden, S. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* **118** (2021).
- 35 Laursen, D. Please reply! The replying norm in adolescent SMS communication. in *The Inside text* 53-73 (Springer, 2005).
- 36 Sousa, D., Sarmiento, L. & Mendes Rodrigues, E. Characterization of the Twitter @replies network: Are user ties social or topical? in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. 63-70.
- 37 Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374-378 (2019).
- 38 Guess, A., Nagler, J., & Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* **5** (2019).
- 39 Guay, B., Pennycook, G., Rand, D. G. Examining partisan asymmetries in fake news sharing and the efficacy of accuracy prompt interventions. *PsyArXiv*, <<https://psyarxiv.com/y762k/>> (2022).
- 40 Munger, K. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* **39**, 629-649 (2017).
- 41 Freelon, D. *et al.* Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*, 0894439320914853 (2020).

- 42 Friggeri, A., Adamic, L., Eckles, D. & Cheng, J. in *Proceedings of the International AAAI*
 43 *Conference on Web and Social Media*.
- 44 Shin, J., Jian, L., Driscoll, K. & Bar, F. Political rumoring on Twitter during the 2012 US
 45 presidential election: Rumor diffusion and correction. *New Media & Society* **19**, 1214-1235
 46 (2017).
- 47 Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. Tweeting from left to right:
 48 Is online political communication more than an echo chamber? *Psychological Science* **26**,
 1531-1542 (2015).
- 49 Higgins, M. J., Sävje, F. & Sekhon, J. S. Improving massive experiments with threshold
 blocking. *Proceedings of the National Academy of Sciences* **113**, 7369-7376 (2016).
- 46 Gilens, N. & Williams, J. *Federal Judge Rules It Is Not a Crime to Violate a Website's*
Terms of Service, <[https://www.eff.org/deeplinks/2020/04/federal-judge-rules-it-not-](https://www.eff.org/deeplinks/2020/04/federal-judge-rules-it-not-crime-violate-websites-terms-service)
[crime-violate-websites-terms-service](https://www.eff.org/deeplinks/2020/04/federal-judge-rules-it-not-crime-violate-websites-terms-service)> (2020).
- 47 Baar, S. V. *Memorandum Opinion*, <[https://www.aclu.org/sandvig-v-barr-memorandum-](https://www.aclu.org/sandvig-v-barr-memorandum-opinion)
[opinion](https://www.aclu.org/sandvig-v-barr-memorandum-opinion)> (2020).
- 48 Cicchetti, D. V. & Sparrow, S. A. Developing criteria for establishing interrater reliability
 of specific items: applications to assessment of adaptive behavior. *American Journal of*
Mental Deficiency (1981).
- 49 Faul, F., Erdfelder, E., Lang, A. & Buchner, A. A flexible statistical power analysis
 program for the social, behavioral and biomedical sciences. *Behavior Research Methods*.

Supplemental Information for:
Promoting engagement with social fact-checks online
Mohsen Mosleh*^{1,2}, Cameron Martel*², Dean Eckles^{2,3}, David G. Rand^{2,3,4}

Corresponding author: mmosleh@mit.edu

*Authors contributed equally

¹SITE (Science, Innovation, Technology, and Entrepreneurship) Department, University of Exeter Business School, Exeter EX4 4PU, UK; ²Sloan School of Management, Massachusetts Institute of Technology, 02142 Cambridge, MA, USA; ³Institute for Data, Systems, and Society, Massachusetts Institute of Technology, 02142 Cambridge, MA, USA; ⁴Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 02139 Cambridge, MA, USA

1. Field Experiment

1.1 Materials

Supplementary Table 1. Full list of 11 political articles rated false by fact-checking website Snopes.com. URLs used to find users who posted links to these articles on Twitter. Articles were selected via collecting the most recent political fact-checks written by Snopes.com in the two months prior to the Twitter field experiment (11-November-2019 to 3-January-2020). Users' partisanship is calculated using¹ on continuous interval $[-2.5, 2.5]$ where -2.5 represents strong liberal and 2.5 represents strong conservative,

	Claim	Fact-check URL	Averaged est. users' partisanship
1	Ukraine donated more money than any other country to the Clinton Foundation.	https://www.snopes.com/fact-check/ukraine-clinton-foundation/	1.61
2	In 2019, a U.S. District Court judge ruled that girls in an Illinois school district "must shower with boys" and had no right to privacy.	https://www.snopes.com/fact-check/girls-shower-with-boys-ruling/	1.19
3	Donald Trump once evicted a disabled combat veteran for owning a small therapy dog.	https://www.snopes.com/fact-check/trump-veteran-service-dog/	-1.09
4	A photograph of U.S. President Donald Trump in his Trump Tower office in 2016 with several boxes of Sudafed in the background provides credible evidence of stimulant abuse.	https://www.snopes.com/fact-check/sudafed-trump-tower/	-0.791
5	The New York Times stated, as fact, that Hillary Clinton and George Soros had been responsible for paying a woman to make false allegations of sexual assault against Donald Trump.	https://www.snopes.com/fact-check/soros-clinton-trump-accuser/	1.71
6	An American diplomat named Melanie Honcharenko was found dead shortly before testifying in the impeachment inquiry against U.S. President Donald Trump.	https://www.snopes.com/fact-check/trump-impeachment-witness-die/	1.64
7	A proposed Virginia law would outlaw martial arts and firearms instruction.	https://www.snopes.com/fact-check/virginia-martial-arts-firearms-law/	1.72
8	"Illegal immigrants" killed 10,150 Americans in 2018.	https://www.snopes.com/fact-check/immigrant-homicides-us/	1.34
9	Virginia Gov. Ralph Northam said the National Guard would cut power and communications before killing anyone who didn't comply with new gun legislation.	https://www.snopes.com/fact-check/northam-meme/	1.74
10	A photograph shows Melania Trump with porn star Ron Jeremy.	https://www.snopes.com/fact-check/melania-trump-ron-jeremy-photo/	-0.265
11	Eric Trump tweeted about the airstrike that killed Iran Gen. Qassem Soleimani in early 2020 before the military operation took place.	https://www.snopes.com/fact-check/eric-trump-tweet-iran-strike/	-1.10

Supplementary Table 2. Full list of variations of the corrective messages used to debunk false content shared by users.

1	I don't know if I believe this article - it might not be true I found a Snopes link that says this headline is false
2	I don't know if I trust this article - it might not be true I found a Snopes link that says this headline is false
3	I'm not sure about the article you posted - it might not be true I found a link on Snopes that says this headline is false
4	I'm not sure about this article - it could be false I found a link on Snopes that says this headline is false
5	I'm not sure about this article - it may not be true I found a link on Snopes that says this headline is false
6	I'm not sure about this article - it might be false I found a link on Snopes that says this headline is false
7	I'm not sure about this article - it might be false I found a link on Snopes that says this headline is not true
8	I'm not sure about this article - it might not be true I found a link on Snopes that says this headline is false
9	I'm not sure about this article - it might not be true I found a link on Snopes that says this headline is not true
10	I'm not sure about this article - it might not be true I found a Snopes link that says this headline is false
11	I'm uncertain about this article - it might not be true I found a link on Snopes that says this headline is false
12	I'm unsure about this article - it might not be true I found a link on Snopes that says this headline is false

1.2 Pre-registered Analyses

In the pre-registration (https://aspredicted.org/F4C_3Y5), we specified our main analysis as a regression model predicting engagement by social condition (-0.5 = baseline, 0.5 = social), bot partisanship (i.e., -0.5=Democrat, 0.5=Republican), participant partisanship (Democrat versus Republican; z-score), and their interactions. In the main text we reported the results using shared partisanship between the bot and the user to be consistent with the analysis for the survey experiment. Here we report results using bot partisanship.

Supplementary Table 3. Descriptive of engagements, positive responses, and negative responses by conditions (N=1,586).

Experimental condition	N_{Total}	$N_{DeliverySuccess}$	Number of engagements	Number of positive responses	Number of negative responses
Baseline counter-partisan	400	382	89	44	45
Social counter-partisan	401	329	89	42	47
Baseline co-partisan	396	381	90	44	46
Social co-partisan	389	364	110	47	63

Supplementary Table 4. Effect of bot partisanship, prior social connection, and user partisanship on engagement.

	b	Std	t	p_{Reg}	p_{FRI}
(Intercept)	0.261	0.012	22.596	<0.001***	<0.001***
Social condition	0.051	0.023	2.221	0.027*	0.027*
Bot partisanship	-0.008	0.012	-0.665	0.506	0.496
User partisanship	0.001	0.012	0.086	0.932	0.8
Social condition $\times$ bot partisanship	0.019	0.023	0.837	0.403	0.4
Social condition $\times$ user partisanship	0.000	0.023	0.011	0.991	0.991
Bot partisanship $\times$ user partisanship	0.015	0.012	1.273	0.203	0.191
Social Condition $\times$ bot partisanship $\times$ user partisanship	0.007	0.023	0.297	0.766	0.744

Supplementary Table 5. Effect of bot partisanship, prior social connection, and user partisanship on positive engagement.

	b	Std	t	p_{Reg}	p_{FRI}
(Intercept)	0.122	0.009	14.167	<0.001***	<0.001***
Social condition	0.013	0.017	0.781	0.436	0.42
Bot partisanship	-0.013	0.009	-1.528	0.127	0.149
User partisanship	-0.017	0.009	-1.937	0.054	0.982
Social condition $\times$ bot partisanship	0.000	0.017	-0.009	0.993	0.99
Social condition $\times$ user partisanship	-0.005	0.017	-0.300	0.764	0.772
Bot partisanship $\times$ user partisanship	0.008	0.009	0.933	0.351	0.387
Social Condition $\times$ bot partisanship $\times$ user partisanship	0.001	0.017	0.051	0.952	0.961

Supplementary Table 6. Effect of bot partisanship, prior social connection, and user partisanship on negative engagement.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p_{Reg}</i>	<i>p_{FRI}</i>
(Intercept)	0.139	0.009	15.301	<0.001***	<0.001***
Social condition	0.038	0.018	2.083	0.037*	0.034*
Bot partisanship	0.005	0.009	0.599	0.549	0.535
User partisanship	0.018	0.009	1.923	0.055	0.997
Social condition × bot partisanship	0.019	0.018	1.033	0.302	0.291
Social condition × user partisanship	0.005	0.018	0.296	0.767	0.743
Bot partisanship × user partisanship	0.007	0.009	0.733	0.463	0.437
Social Condition × bot partisanship × user partisanship	0.006	0.018	0.307	0.759	0.764

Supplementary Figure 1. Effect of bot partisanship, prior social connection, and user partisanship on engagement.**Supplementary Table 7.** Significance of difference between coefficients of positive engagement and negative engagement.

	<i>p_{Reg}</i>	<i>p_{FRI}</i>
(Intercept)	0.176	0.389
Social condition	0.334	0.349
Shared partisanship	0.137	0.155
User partisanship	0.006	0.157
Social condition × Bot partisanship	0.456	0.492
Social condition × User partisanship	0.685	0.680
Bot partisanship × User partisanship	0.921	0.890
Social Condition × Shared partisanship × User partisanship	0.859	0.876

1.3 Full Main Analyses

Supplementary Table 8. Effect of shared partisanship, prior social connection, and user partisanship on engagement ($N=1,456$).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i> _{Reg}	<i>p</i> _{FRI}
(Intercept)	0.261	0.012	22.618	<0.001***	<0.001***
Social condition	0.052	0.023	2.242	0.025*	0.017*
Shared partisanship	0.009	0.012	0.748	0.454	0.435
User partisanship	0.001	0.012	0.064	0.949	0.947
Social condition × Shared partisanship	0.016	0.023	0.690	0.490	0.475
Social condition × User partisanship	0.001	0.023	0.026	0.979	0.974
Shared partisanship × User partisanship	-0.014	0.012	-1.227	0.220	0.200
Social Condition × Shared partisanship × User partisanship	0.013	0.023	0.556	0.578	0.568

Supplementary Table 9. Effect of shared partisanship, prior social connection, and user partisanship on positive engagement ($N=1,456$).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i> _{Reg}	<i>p</i> _{FRI}
(Intercept)	0.122	0.009	14.164	<0.001***	<0.001***
Social condition	0.013	0.017	0.782	0.435	0.430
Shared partisanship	0.000	0.009	0.006	0.995	0.996
User partisanship	-0.017	0.009	-1.975	0.050*	0.747
Social condition × Shared partisanship	0.001	0.017	0.039	0.955	0.961
Social condition × User partisanship	-0.005	0.017	-0.301	0.764	0.744
Shared partisanship × User partisanship	-0.015	0.009	-1.783	0.075	0.110
Social Condition × Shared partisanship × User partisanship	-0.001	0.017	-0.034	0.981	0.973

Supplementary Table 10. Effect of shared partisanship, prior social connection, and user partisanship on negative engagement ($N=1,456$).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i> _{Reg}	<i>p</i> _{FRI}
(Intercept)	0.139	0.009	15.336	<0.001***	<0.001***
Social condition	0.038	0.018	2.111	0.035*	0.033*
Shared partisanship	0.009	0.009	0.945	0.345	0.303
User partisanship	0.018	0.009	1.944	0.052	0.989
Social condition × Shared partisanship	0.015	0.018	0.824	0.41	0.41
Social condition × User partisanship	0.006	0.018	0.318	0.751	0.74
Shared partisanship × User partisanship	0.001	0.009	0.130	0.897	0.883

Social Condition $\times$ Shared partisanship $\times$ User partisanship	0.013	0.018	0.730	0.465	0.432
--	-------	-------	-------	-------	-------

Supplementary Table 11. Significance of difference between coefficients of positive engagement and negative engagement.

	p_{Reg}	p_{FRI}
(Intercept)	0.168	0.365
Social condition	0.325	0.395
Shared partisanship	0.506	0.564
User partisanship	0.006	0.150
Social condition $\times$ Shared partisanship	0.588	0.629
Social condition $\times$ User partisanship	0.675	0.706
Shared partisanship $\times$ User partisanship	0.188	0.203
Social Condition $\times$ Shared partisanship $\times$ User partisanship	0.597	0.620

Supplementary Figure 2. Effect of shared partisanship, user partisanship, and prior social connection on probability of engagement.

1.4 Bayesian analyses

Supplementary Table 12. Bayesian regression analysis of effect of shared partisanship, prior social connection, and user partisanship on engagement using weakly-informative normal prior over intercept and effect of predictors and weakly-informative prior over residual. l-95% CI shows the 0.025 percentile of posterior distribution and u-95% CI shows the 0.075 percentile of posterior distribution.

	Estimate	Estimate err.	l-95%	u-95%	$\hat{R}$	Bulk ESS	Tail ESS
(Intercept)	0.26	0.01	0.24	0.28	1	16339	6473
Social condition	0.05	0.02	0.01	0.1	1	14461	6193
Shared partisanship	0.01	0.01	-0.01	0.03	1	13661	6231
User partisanship	0	0.01	-0.02	0.02	1	14880	5444
Social condition $\times$ Shared partisanship	0.02	0.02	-0.03	0.06	1	15653	6094
Social condition $\times$ User partisanship	0	0.02	-0.05	0.05	1	13604	5975
Shared partisanship $\times$ User partisanship	-0.01	0.01	-0.04	0.01	1	15634	5965
Social Condition $\times$ Shared partisanship $\times$ User partisanship	0.01	0.02	-0.03	0.06	1	17447	6649

Supplementary Table 13. Bayesian regression analysis of effect of shared partisanship, prior social connection, and user partisanship on positive engagement using weakly-informative normal prior over intercept and effect of predictors and weakly-informative prior over residual. l-95% CI shows the 0.025 percentile of posterior distribution and u-95% CI shows the 0.075 percentile of posterior distribution.

	Estimate	Estimate err.	l-95%	u-95%	$\hat{R}$	Bulk ESS	Tail ESS
(Intercept)	0.12	0.01	0.1	0.14	1	14595	6123
Social condition	0.01	0.02	-0.02	0.05	1	14362	6170
Shared partisanship	0	0.01	-0.02	0.02	1	14694	5437
User partisanship	-0.02	0.01	-0.03	0	1	14133	6570
Social condition $\times$ Shared partisanship	0	0.02	-0.03	0.04	1	14967	5848
Social condition $\times$ User partisanship	0	0.02	-0.04	0.03	1	13998	6357
Shared partisanship $\times$ User partisanship	-0.02	0.01	-0.03	0	1	14360	6119
Social Condition $\times$ Shared partisanship $\times$ User partisanship	0	0.02	-0.03	0.03	1	14355	6038

Supplementary Table 14. Bayesian regression analysis of effect of shared partisanship, prior social connection, and user partisanship on negative engagement using weakly-informative normal prior over intercept and effect of predictors and weakly-informative prior over residual. l-95% CI shows the 0.025 percentile of posterior distribution and u-95% CI shows the 0.075 percentile of posterior distribution.

	Estimate	Estimate err.	l-95%	u-95%	$\hat{R}$	Bulk ESS	Tail ESS
(Intercept)	0.14	0.01	0.12	0.16	1	14265	5506
Social condition	0.04	0.02	0	0.07	1	14102	6267
Shared partisanship	0.01	0.01	-0.01	0.03	1	13703	5730
User partisanship	0.02	0.01	0	0.04	1	13763	6664
Social condition $\times$ Shared partisanship	0.01	0.02	-0.02	0.05	1	14073	5974
Social condition $\times$ User partisanship	0.01	0.02	-0.03	0.04	1	13206	6219
Shared partisanship $\times$ User partisanship	0	0.01	-0.02	0.02	1	14021	6609
Social Condition $\times$ Shared partisanship $\times$ User partisanship	0.01	0.02	-0.02	0.05	1	14857	6016

1.5 Article-level and corrective message level analysis

Supplementary Figure 3. Probability of user engagement with corrective message for each debunked article. The number of observations per each article is not sufficient to draw meaningful statistical inference in terms of heterogeneous effect of the article on probability of engagement. Article numbers correspond to Supplementary Table 1.

Supplementary Figure 4. Probability of user engagement with corrective message for each correction message variation. Corrective message wordings were not fully randomized as part of our experimental procedure. Message numbers correspond to Supplementary Table 2.

1.6 Reply Tweet Rating

In order to assess whether reply tweets in the field experiment were receptive or rejecting of our corrective messages, we recruited 66 Amazon Mechanical Turk workers to each evaluate 50 unique reply tweets and rate them as either ‘belief updating’ or ‘resisting information.’ Raters were told they would be asked to read a number of tweets, where in each case the writer had first shared a news article on Twitter, and then had received a response message indicating that the news article was false. Raters were also told that the correction message included a real link to a Snopes article. Raters then were instructed that they would be reading the replies Twitter users then wrote to such correction messages. For each reply, raters were asked to categorize the reply as either ‘belief updating’ or ‘resisting information,’ defined as follows:

Belief Updating: Replies that indicate that the replier originally thought the article they shared was true, but after seeing the correction message now believe that what they shared is false. Belief updating can be thought of as the correction successfully changing the mind of the replier, such that they now know they shared fake news. These replies may involve saying **thank you**, saying **sorry**, or indicating that they will **delete their original post**.

Example - “Thank you for your message, I guess I was wrong and the article I posted was fake.”

Example - “Thanks, sorry I usually fact check my posts better. I will delete my Tweet.”

Resisting Information: Replies that argue that their original post is true, argue that the correction itself is false or misleading, or argue in support of the general idea or sentiment behind the original post being true. Generally speaking, any attempt to counter-argue or ignore the correction is an example of resisting information. Replies that do not engage with the correction and simply indicate some level of ignorance or apathy towards the original poster also constitute resisting information. These replies may involve saying that the **correction is fake**, arguing that **there is information refuting the correction**, or **ignoring the correction** while making other arguments or claims.

Example - “I think the article I posted is correct, because I’ve read it on other news sites.”

Example - “Snopes is fake news.”

Example - “I don’t know what to believe anymore.”

Raters were also instructed that all Twitter handles had been anonymized to “@username,” but that other than this modification, all tweets were presented as they actually appeared on Twitter.

If a user in the field experiment replied to a correction message with multiple tweets, these tweets were combined into a single unique tweet reply for rating purposes.

For each reply tweet, raters first read the tweet, and then were asked: “Does this reply belong to the category of **belief updating** (e.g., “Thanks, sorry, I usually fact check my posts better. I will delete by Tweet.”) or **resisting information** (e.g., “I think the article I posted is correct, because I’ve read it on other news sites. Also, Snopes is fake news.”)? (*Belief Updating; Resisting Information*).

For each reply tweet, the aggregate reply tweet type was recorded as the modal rating. Deleted tweets and favourited or retweeted correction messages (if there was no written reply) were also recorded as types of positive engagement. Resisting information was recorded as negative engagement.

Supplementary Table 15. Different types of tweet reactions coded as positive engagement.

Types of Positive Engagement		
Modal rater evaluation of written reply indicated response expressed belief updating (as defined above)	Original tweet containing false news was deleted by user after successful correction delivery	User favourited or retweeted the correction message (conditional on there being no written reply)
136	13	28

1.7 Follow-back Analysis

Supplementary Table 16. Effect of shared partisanship and user partisanship on probability of follow-back in social condition.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.230	0.016	14.807	<0.001***
Shared partisanship	0.094	0.016	6.086	<0.001***
User partisanship	0.060	0.016	3.837	<0.001***
Shared partisanship × User partisanship	-0.003	0.016	-0.165	0.869

Supplementary Figure 5. Effect of shared partisanship and user partisanship on probability of follow-back in social condition.

Supplementary Table 17. Effect of bot partisanship and user partisanship on probability of follow-back in social condition.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.224	0.016	14.445	<0.001***
Bot partisanship	0.050	0.016	3.204	0.001**
User partisanship	0.059	0.016	3.759	<0.001***
Bot partisanship $\times$ user partisanship	0.081	0.016	5.150	<0.001***

Supplementary Figure 6. Effect of bot partisanship and user partisanship on probability of follow-back in social condition.

2. Survey Experiment 1

2.1. Pre-registration.

In our pre-registration (https://aspredicted.org/ZHT_SQW), we specified our main analysis as a regression model predicting our key dependent variables (engagement; various mechanism questions) by social condition (-0.5 = baseline, 0.5 = social), political concordance (i.e., shared partisanship; -0.5=counter-partisan, 0.5=co-partisan), and their interaction. As our main reported analyses, we deviate from this pre-registration plan to also include a z-scored continuous measure of partisanship as a predictor, allowing for all interactions. However, we report full analyses and results here for these originally pre-registered analyses.

2.2. Descriptive Statistics.

Supplementary Table 18. Reply engagement descriptive statistics by condition. Total N=808.

Baseline, Counter-partisan	Baseline, Co-partisan	Social, Counter-partisan	Social, Co-partisan
Engaged: 102 Ignored: 100 M _{engage} : 50.50%	Engaged: 105 Ignored: 99 M _{engage} : 51.47%	Engaged: 107 Ignored: 89 M _{engage} : 54.59%	Engaged: 133 Ignored: 73 M _{engage} : 64.56%

2.3. Main Analyses.

Supplementary Table 19. Reply engagement (binary) predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.553	0.017	31.828	<0.001***
Social condition	0.083	0.035	2.373	0.018*
Participant partisanship	-0.048	0.017	-2.739	0.006**
Shared partisanship	0.028	0.017	1.608	0.108
Social condition × Participant partisanship	0.001	0.035	0.039	0.969
Social condition × Shared partisanship	0.043	0.035	1.242	0.214
Participant partisanship × Shared partisanship	-0.012	0.017	-0.699	0.485
Social condition × Participant partisanship × Shared partisanship	0.024	0.035	0.687	0.492

Supplementary Table 20. Rudeness of non-responding predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
--	----------	------------	----------	----------

(Intercept)	3.296	0.072	45.857	<0.001***
Social condition	0.174	0.144	1.208	0.227
Participant partisanship	-0.152	0.072	-2.101	0.036*
Shared partisanship	0.019	0.072	0.261	0.794
Social condition × Participant partisanship	-0.064	0.144	-0.445	0.656
Social condition × Shared partisanship	0.067	0.144	0.467	0.640
Participant partisanship × Shared partisanship	-0.062	0.072	-0.859	0.391
Social condition × Participant partisanship × Shared partisanship	-0.015	0.144	-0.107	0.915

Supplementary Table 21. *Obligation to respond predicted by social condition, participant partisanship, shared partisanship, and all interactions.*

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.751	0.074	50.882	<0.001***
Social condition	0.3	0.147	2.031	0.043*
Participant partisanship	-0.142	0.074	-1.921	0.055
Shared partisanship	0.005	0.074	0.074	0.941
Social condition × Participant partisanship	-0.015	0.148	-0.104	0.917
Social condition × Shared partisanship	0.07	0.148	0.476	0.634
Participant partisanship × Shared partisanship	-0.056	0.074	-0.759	0.448
Social condition × Participant partisanship × Shared partisanship	-0.042	0.148	-0.282	0.778

Supplementary Table 22. *Appropriateness of correcting others predicted by social condition, participant partisanship, shared partisanship, and all interactions.*

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	5.168	0.041	125.694	<0.001***
Social condition	0.074	0.082	0.899	0.369
Participant partisanship	-0.169	0.041	-4.096	<0.001***
Shared partisanship	0.022	0.041	0.544	0.587
Social condition × Participant partisanship	0.098	0.082	1.182	0.237
Social condition × Shared partisanship	-0.02	0.082	-0.243	0.808
Participant partisanship × Shared partisanship	-0.002	0.041	-0.054	0.957
Social condition × Participant partisanship × Shared partisanship	0.174	0.083	2.108	0.035*

Supplementary Table 23. Comfort with corrector (reverse-coded) predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.798	0.06	62.809	<0.001***
Social condition	-0.031	0.121	-0.254	0.799
Participant partisanship	-0.095	0.061	-1.571	0.117
Shared partisanship	-0.113	0.061	-1.862	0.063
Social condition × Participant partisanship	-0.051	0.121	-0.419	0.675
Social condition × Shared partisanship	-0.124	0.121	-1.024	0.306
Participant partisanship × Shared partisanship	-0.053	0.061	-0.869	0.385
Social condition × Participant partisanship × Shared partisanship	0.111	0.121	0.91	0.363

Supplementary Table 24. Trustworthiness of corrector predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.705	0.048	98.269	<0.001***
Social condition	0.092	0.096	0.962	0.336
Participant partisanship	-0.13	0.048	-2.7	0.007**
Shared partisanship	0.117	0.048	2.444	0.015*
Social condition × Participant partisanship	0.089	0.096	0.928	0.354
Social condition × Shared partisanship	0.043	0.096	0.446	0.656
Participant partisanship × Shared partisanship	-0.062	0.048	-1.3	0.194
Social condition × Participant partisanship × Shared partisanship	0.048	0.096	0.494	0.621

Supplementary Table 25. How smart is corrector predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.988	0.045	112.044	<0.001***
Social condition	0.132	0.089	1.486	0.138
Participant partisanship	-0.125	0.045	-2.8	0.005**
Shared partisanship	0.114	0.045	2.568	0.010**
Social condition × Participant partisanship	0.059	0.089	0.663	0.508

Social condition × Shared partisanship	0.014	0.089	0.159	0.874
Participant partisanship × Shared partisanship	-0.066	0.045	-1.467	0.143
Social condition × Participant partisanship × Shared partisanship	0.082	0.089	0.922	0.357

Supplementary Table 26. Positive/negative feelings towards corrector predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.69	0.048	98.01	<0.001***
Social condition	0.141	0.096	1.473	0.141
Participant partisanship	-0.12	0.048	-2.505	0.012*
Shared partisanship	0.154	0.048	3.216	0.001***
Social condition × Participant partisanship	0.109	0.096	1.138	0.255
Social condition × Shared partisanship	0.115	0.096	1.196	0.232
Participant partisanship × Shared partisanship	-0.096	0.048	-1.987	0.047*
Social condition × Participant partisanship × Shared partisanship	0.023	0.096	0.234	0.815

Supplementary Table 27. Likelihood corrector is bot predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.427	0.06	73.483	<0.001***
Social condition	-0.002	0.12	-0.014	0.989
Participant partisanship	-0.192	0.06	-3.182	0.002**
Shared partisanship	-0.028	0.06	-0.465	0.642
Social condition × Participant partisanship	-0.094	0.121	-0.775	0.439
Social condition × Shared partisanship	-0.072	0.121	-0.595	0.552
Participant partisanship × Shared partisanship	0.038	0.061	0.628	0.530
Social condition × Participant partisanship × Shared partisanship	-0.039	0.121	-0.326	0.744

Supplementary Table 28. Obligated self-correction (exploratory) predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.688	0.062	75.578	<0.001***

Social condition	0.066	0.124	0.528	0.597
Participant partisanship	-0.239	0.062	-3.834	<0.001***
Shared partisanship	0.043	0.062	0.686	0.493
Social condition × Participant partisanship	0.082	0.124	0.658	0.511
Social condition × Shared partisanship	0.009	0.124	0.073	0.942
Participant partisanship × Shared partisanship	-0.026	0.062	-0.421	0.674
Social condition × Participant partisanship × Shared partisanship	0.014	0.125	0.114	0.910

2.4. Main Analyses – Binary Candidate Choice Partisanship.

In addition to primarily assessing partisanship via a 1-7 Likert-scale, we also included a binary choice Presidential candidate measure of partisanship. Here, we repeat our main analyses except replacing our continuous measure of partisanship with this binary choice measure of partisanship. Crucially, we again observe a significant main effect of prior social connection increasing likelihood of responding to a correction message; we again do not observe an effect of shared partisanship on reply likelihood, nor do we observe an interaction between prior social connection and shared partisanship on reply likelihood.

Supplementary Table 29. Reply engagement (binary) predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.553	0.017	32.265	<0.001***
Social condition	0.089	0.034	2.604	0.009**
Participant partisanship	0.091	0.017	5.330	<0.001***
Shared partisanship	0.015	0.017	0.882	0.378
Social condition × Participant partisanship	-0.018	0.034	-0.524	0.600
Social condition × Shared partisanship	0.028	0.034	0.812	0.417
Participant partisanship × Shared partisanship	-0.034	0.017	-1.982	0.048*
Social condition × Participant partisanship × Shared partisanship	-0.006	0.034	-0.168	0.866

Supplementary Table 30. Rudeness of non-responding predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.297	0.067	49.133	<0.001***
Social condition	0.200	0.134	1.494	0.136

Participant partisanship	0.740	0.067	11.025	<0.001***
Shared partisanship	-0.011	0.067	-0.169	0.866
Social condition × Participant partisanship	-0.149	0.134	-1.111	0.267
Social condition × Shared partisanship	-0.007	0.134	-0.054	0.957
Participant partisanship × Shared partisanship	-0.043	0.067	-0.641	0.521
Social condition × Participant partisanship × Shared partisanship	-0.040	0.134	-0.297	0.766

Supplementary Table 31. *Obligation to respond predicted by social condition, participant partisanship, shared partisanship, and all interactions.*

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.753	0.068	54.86	<0.001***
Social condition	0.329	0.137	2.405	0.016*
Participant partisanship	0.784	0.068	11.445	<0.001***
Shared partisanship	-0.001	0.068	-0.014	0.988
Social condition × Participant partisanship	0.029	0.137	0.211	0.833
Social condition × Shared partisanship	0.018	0.137	0.130	0.897
Participant partisanship × Shared partisanship	-0.072	0.068	-1.054	0.292
Social condition × Participant partisanship × Shared partisanship	-0.140	0.137	-1.024	0.306

Supplementary Table 32. *Appropriateness of correcting others predicted by social condition, participant partisanship, shared partisanship, and all interactions.*

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	5.165	0.042	124.195	<0.001***
Social condition	0.088	0.083	1.061	0.289
Participant partisanship	0.071	0.042	1.701	0.089
Shared partisanship	0.037	0.042	0.893	0.372
Social condition × Participant partisanship	0.063	0.083	0.753	0.451
Social condition × Shared partisanship	0.016	0.083	0.189	0.850
Participant partisanship × Shared partisanship	0.004	0.042	0.104	0.917
Social condition × Participant partisanship × Shared partisanship	0.029	0.083	0.347	0.729

Supplementary Table 33. Comfort with corrector (reverse-coded) predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.798	0.061	62.71	<0.001***
Social condition	-0.023	0.121	-0.186	0.853
Participant partisanship	-0.043	0.061	-0.706	0.480
Shared partisanship	-0.112	0.061	-1.848	0.065
Social condition × Participant partisanship	0.015	0.121	0.124	0.902
Social condition × Shared partisanship	-0.076	0.121	-0.625	0.532
Participant partisanship × Shared partisanship	0.005	0.061	0.088	0.930
Social condition × Participant partisanship × Shared partisanship	0.128	0.121	1.058	0.290

Supplementary Table 34. Trustworthiness of corrector predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.703	0.047	100.515	<0.001***
Social condition	0.109	0.094	1.170	0.242
Participant partisanship	0.268	0.047	5.720	<0.001***
Shared partisanship	0.156	0.047	3.340	0.001***
Social condition × Participant partisanship	0.063	0.094	0.678	0.498
Social condition × Shared partisanship	0.149	0.094	1.596	0.111
Participant partisanship × Shared partisanship	-0.123	0.047	-2.632	0.009**
Social condition × Participant partisanship × Shared partisanship	0.008	0.094	0.083	0.934

Supplementary Table 35. How smart is corrector predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.987	0.043	114.863	<0.001***
Social condition	0.151	0.087	1.735	0.083
Participant partisanship	0.280	0.043	6.447	<0.001***
Shared partisanship	0.115	0.043	2.647	0.008**
Social condition × Participant partisanship	0.093	0.087	1.068	0.286

Social condition × Shared partisanship	0.063	0.087	0.729	0.466
Participant partisanship × Shared partisanship	-0.126	0.043	-2.899	0.004**
Social condition × Participant partisanship × Shared partisanship	0.062	0.087	0.719	0.473

Supplementary Table 36. Positive/negative feelings towards corrector predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.689	0.046	101.933	<0.001***
Social condition	0.162	0.092	1.761	0.079
Participant partisanship	0.353	0.046	7.671	<0.001***
Shared partisanship	0.157	0.046	3.414	0.001***
Social condition × Participant partisanship	0.143	0.092	1.549	0.122
Social condition × Shared partisanship	0.161	0.092	1.752	0.080
Participant partisanship × Shared partisanship	-0.169	0.046	-3.676	<0.001***
Social condition × Participant partisanship × Shared partisanship	-0.076	0.092	-0.829	0.408

Supplementary Table 37. Likelihood corrector is bot predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.430	0.058	76.405	<0.001***
Social condition	0.020	0.116	0.171	0.864
Participant partisanship	0.492	0.058	8.476	<0.001***
Shared partisanship	-0.025	0.058	-0.425	0.671
Social condition × Participant partisanship	0.131	0.116	1.132	0.258
Social condition × Shared partisanship	-0.027	0.116	-0.233	0.816
Participant partisanship × Shared partisanship	0.077	0.058	1.321	0.187
Social condition × Participant partisanship × Shared partisanship	0.031	0.116	0.268	0.789

Supplementary Table 38. Obligated self-correction (exploratory) predicted by social condition, participant partisanship, shared partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
--	----------	------------	----------	----------

(Intercept)	4.686	0.062	75.216	<0.001***
Social condition	0.090	0.125	0.722	0.471
Participant partisanship	0.144	0.062	2.317	0.021*
Shared partisanship	0.033	0.062	0.530	0.596
Social condition × Participant partisanship	-0.022	0.125	-0.180	0.857
Social condition × Shared partisanship	0.059	0.125	0.476	0.635
Participant partisanship × Shared partisanship	-0.066	0.062	-1.057	0.291
Social condition × Participant partisanship × Shared partisanship	-0.113	0.125	-0.906	0.365

2.5. Pre-registered Analyses.

Supplementary Table 39. Reply engagement (binary) predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.553	0.017	31.752	<0.001***
Social condition	0.087	0.035	2.486	0.013*
Shared partisanship	0.027	0.017	1.571	0.117
Social condition × Shared partisanship	0.045	0.035	1.291	0.197

Supplementary Table 40. Rudeness of non-responding predicted by social condition, shared partisanship, and their interaction

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.298	0.072	45.872	<0.001***
Social condition	0.187	0.144	1.298	0.195
Shared partisanship	0.019	0.072	0.267	0.790
Social condition × Shared partisanship	0.075	0.144	0.521	0.603

Supplementary Table 41. Obligation to respond predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.752	0.074	50.919	<0.001***
Social condition	0.312	0.147	2.116	0.035*
Shared partisanship	0.006	0.074	0.082	0.935
Social condition × Shared partisanship	0.077	0.147	0.523	0.601

Supplementary Table 42. Appropriateness of correcting others predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	5.165	0.042	124.205	<0.001***
Social condition	0.088	0.083	1.057	0.291
Shared partisanship	0.018	0.042	0.436	0.663
Social condition × Shared partisanship	-0.017	0.083	-0.202	0.840

Supplementary Table 43. Comfort with corrector (reverse-coded) predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.799	0.06	62.855	<0.001***
Social condition	-0.022	0.121	-0.182	0.856
Shared partisanship	-0.115	0.06	-1.899	0.058
Social condition × Shared partisanship	-0.118	0.121	-0.978	0.329

Supplementary Table 44. Trustworthiness of corrector predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.703	0.048	97.893	<0.001***
Social condition	0.104	0.096	1.079	0.281
Shared partisanship	0.115	0.048	2.403	0.017*
Social condition × Shared partisanship	0.05	0.096	0.521	0.602

Supplementary Table 45. How smart is corrector predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.987	0.045	111.539	<0.001***
Social condition	0.144	0.089	1.607	0.108
Shared partisanship	0.112	0.045	2.509	0.012*
Social condition × Shared partisanship	0.022	0.089	0.241	0.810

Supplementary Table 46. Positive/negative feelings towards corrector predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.688	0.048	97.531	<0.001***
Social condition	0.152	0.096	1.584	0.114
Shared partisanship	0.153	0.048	3.178	0.002**

Social condition × Shared partisanship	0.124	0.096	1.293	0.196
--	-------	-------	-------	-------

Supplementary Table 47. Likelihood corrector is bot predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.429	0.06	73.266	<0.001***
Social condition	0.014	0.121	0.113	0.910
Shared partisanship	-0.027	0.06	-0.446	0.656
Social condition × Shared partisanship	-0.072	0.121	-0.598	0.550

Supplementary Table 48. Obligated self-correction (exploratory) predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.686	0.062	75.062	<0.001***
Social condition	0.085	0.125	0.682	0.495
Shared partisanship	0.041	0.062	0.663	0.507
Social condition × Shared partisanship	0.015	0.125	0.122	0.903

2.6. Pre-registered Analyses – Binary Candidate Choice Partisanship.

Supplementary Table 49. Reply engagement (binary) predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.553	0.017	31.700	<0.001***
Social condition	0.087	0.035	2.484	0.013*
Shared partisanship	0.016	0.017	0.889	0.374
Social condition × Shared partisanship	0.031	0.035	0.881	0.379

Supplementary Table 50. Rudeness of non-responding predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.30	0.072	45.865	<0.001***
Social condition	0.187	0.144	1.303	0.193
Shared partisanship	-0.009	0.072	-0.129	0.897
Social condition × Shared partisanship	0.010	0.144	0.069	0.945

Supplementary Table 51. Obligation to respond predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
--	----------	------------	----------	----------

(Intercept)	3.75	0.074	50.910	<0.001***
Social condition	0.311	0.147	2.115	0.035*
Shared partisanship	0.004	0.074	0.054	0.957
Social condition × Shared partisanship	0.040	0.147	0.269	0.788

Supplementary Table 52. Appropriateness of correcting others predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	5.165	0.042	124.241	<0.001***
Social condition	0.087	0.083	1.048	0.295
Shared partisanship	0.038	0.042	0.908	0.364
Social condition × Shared partisanship	0.018	0.083	0.214	0.831

Supplementary Table 53. Comfort with corrector (reverse-coded) predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.799	0.06	62.825	<0.001***
Social condition	-0.021	0.121	-0.173	0.863
Shared partisanship	-0.113	0.061	-1.863	0.063
Social condition × Shared partisanship	-0.077	0.121	-0.635	0.525

Supplementary Table 54. Trustworthiness of corrector predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.702	0.048	98.325	<0.001***
Social condition	0.101	0.096	1.059	0.290
Shared partisanship	0.158	0.048	3.305	0.001***
Social condition × Shared partisanship	0.159	0.096	1.662	0.097

Supplementary Table 55. How smart is corrector predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.986	0.045	111.600	<0.001***
Social condition	0.142	0.089	1.594	0.111
Shared partisanship	0.117	0.045	2.612	0.009*
Social condition × Shared partisanship	0.073	0.089	0.821	0.412

Supplementary Table 56. Positive/negative feelings towards corrector predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.687	0.048	97.670	<0.001***
Social condition	0.151	0.096	1.569	0.117
Shared partisanship	0.160	0.048	3.338	0.001***
Social condition × Shared partisanship	0.174	0.096	1.815	0.070

Supplementary Table 57. Likelihood corrector is bot predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.429	0.06	73.236	<0.001***
Social condition	0.014	0.121	0.114	0.909
Shared partisanship	-0.022	0.061	-0.362	0.717
Social condition × Shared partisanship	-0.04	0.121	-0.115	0.908

Supplementary Table 58. Obligated self-correction (exploratory) predicted by social condition, shared partisanship, and their interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.865	0.062	75.056	<0.001***
Social condition	0.085	0.125	0.682	0.496
Shared partisanship	0.034	0.062	0.548	0.584
Social condition × Shared partisanship	0.064	0.125	0.514	0.608

2.7. False Discovery Rate Analyses for Obligated Responding.

Survey experiment 1 provided tentative evidence that prior social connection increased feeling obligated to respond to a correction message. We next correct for multiple comparisons by calculating q-values, which indicate the probability of making at least one false discovery across all comparisons. We first computed a calculated q-value for obligated responding as $1-(1-p)^n$, where n is the number of comparisons for main effects of prior social connection on mechanism (i.e., $n=8$; 8 mechanisms, not including obligated self-correction (exploratory), and $p = .0426$). This calculated q-value is 0.294 (see OSF analysis file https://osf.io/nvk4u/?view_only=2ba8efe34cd647e18f69857bd1f85bbd).

Second, we computed a simulated q-value, which accounts for the fact that the assumption of independence between comparisons may be overly conservative. We simulate the expected probability of making at least one false discovery over 10,000 iterations, following the procedure of ². Each iteration includes randomly sampling observations with replacement, randomly assigning observations to conditions, and conducting all relevant comparison tests. Then simulated q-value is computed as the proportion of iterations in which the minimum simulated p-value is less

than the given p -value. Our simulated q -value was 0.225 (see OSF analysis file https://osf.io/nvk4u/?view_only=2ba8efe34cd647e18f69857bd1f85bbd for full simulation).

Similar false discovery rate analyses were performed to examine the effect of shared partisanship on positive (versus negative) feelings towards the corrector; perceived corrector trustworthiness; and perceived corrector intelligence. Full results reported in the main text, and full analysis code also present in OSF analysis file.

Given that the effect of prior social connection on obligation to respond in our first survey experiment was no longer significant after adjusting for multiple comparisons, we next conducted a follow-up survey experiment to further examine obligation to respond as a mechanism behind the relationship between minimal social relationships and engaging with correction messages.

3. Survey Experiment 2

3.1. Pre-registration.

We pre-registered this study here (https://aspredicted.org/G8Q_413). We made no deviations from this pre-registration.

3.2. Descriptive Statistics.

Supplementary Table 59. Personal obligation to respond descriptive statistics by condition (N=812).

Baseline, Non-correction	Baseline, Correction	Social, Non-correction	Social, Correction
N = 229 M _{obligated} = 3.45 SD _{obligated} = 2.22	N = 191 M _{obligated} = 3.58 SD _{obligated} = 2.08	N = 194 M _{obligated} = 4.12 SD _{obligated} = 2.04	N = 198 M _{obligated} = 4.09 SD _{obligated} = 2.08

Supplementary Table 60. Perceptions of reciprocal responding norm descriptive statistics by condition (N=794).

Baseline, Non-correction	Baseline, Correction	Social, Non-correction	Social, Correction
N = 185 M _{norm} = 4.44 SD _{norm} = 1.94	N = 197 M _{norm} = 4.81 SD _{norm} = 1.65	N = 204 M _{norm} = 4.72 SD _{norm} = 1.67	N = 208 M _{norm} = 5.03 SD _{norm} = 1.56

3.3. Main Analyses.

Supplementary Table 61. Personal obligation to respond predicted by social condition, correction condition, and interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	3.809	0.074	51.285	<0.001***
Social condition	0.592	0.149	3.985	<0.001***
Correction condition	0.049	0.149	0.332	0.740
Social condition × Correction condition	-0.154	0.297	-0.518	0.605

Supplementary Table 62. Perceived norm of responding predicted by social condition, correction condition, and interaction.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.75	0.061	78.415	<0.001***
Social condition	0.245	0.121	2.019	0.044*
Correction condition	0.341	0.121	2.815	0.005**
Social condition × Correction condition	-0.056	0.242	-0.23	0.818

Supplementary Table 63. Personal obligation to respond predicted by social condition, correction condition, partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	4.453	0.141	31.505	<0.001***
Social condition	0.556	0.283	1.965	0.050*
Correction condition	0.322	0.283	1.14	0.254
Participant partisanship	-0.215	0.04	-5.326	<0.001***
Social condition × Correction condition	-0.23	0.565	-0.406	0.684
Social condition × Participant partisanship	-0.005	0.081	-0.065	0.948
Correction condition × Participant partisanship	-0.079	0.081	-0.979	0.328
Social condition × Correction condition × Participant partisanship	0.009	0.162	0.058	0.953

Supplementary Table 64. Perceived norm of responding predicted by social condition, correction condition, partisanship, and all interactions.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	5.07	0.115	44.232	<0.001***
Social condition	0.194	0.229	0.847	0.397
Correction condition	0.117	0.229	0.511	0.609
Participant partisanship	-0.107	0.032	-3.344	0.001***
Social condition × Correction condition	-0.629	0.459	-1.372	0.171
Social condition × Participant partisanship	0.018	0.064	0.285	0.775
Correction condition × Participant partisanship	0.081	0.064	1.253	0.211
Social condition × Correction condition × Participant partisanship	0.182	0.129	1.419	0.156

4. Partisan Extremity Analyses: Field Experiment & Survey Experiment 1

The following exploratory analyses examine the role of partisan extremity (absolute value of z-scored partisanship) predicting correction engagement across social connection and shared partisanship conditions.

4.1. Relationship between Partisan Extremity and Correction Delivery Failure (i.e., Blocking) in Field Experiment.

Supplementary Table 65. Correction delivery failure (i.e., blocking) predicted by user partisan extremity in Social, Counter-partisan field experiment condition.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.035	0.061	0.570	0.569
Social condition	0.093	0.038	2.460	0.014*

Supplementary Table 66. User partisan extremity by field experiment condition

Baseline, Counter-partisan	Baseline, Co-partisan	Social, Counter-partisan	Social, Co-partisan
Extremity: 1.52	Extremity: 1.52	Extremity: 1.50	Extremity: 1.53

Supplementary Table 67. ANOVA predicting user partisan extremity by field experiment condition.

	<i>df</i>	<i>Mean sq.</i>	<i>F</i>	<i>p</i>
Condition	3	0.042	0.16	0.920
Residuals	1452	0.263	---	---

4.2. User Partisan Extremity in Field Experiment

Supplementary Table 665. Effect of shared partisanship, prior social connection, and user partisan extremity on engagement (N=1,456).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.261	0.011	22.796	<0.001***
Social condition	0.053	0.023	2.305	0.021*
Shared partisanship	0.008	0.011	0.710	0.478
User partisan extremity	0.049	0.011	4.269	<0.001***
Social condition × Shared partisanship	0.013	0.023	0.547	0.584
Social condition × User partisan extremity	0.021	0.023	0.921	0.357
Shared partisanship × User partisan extremity	-0.008	0.011	-0.684	0.494
Social Condition × Shared partisanship × User partisan extremity	0.017	0.023	0.755	0.451

Supplementary Table 69. Effect of shared partisanship, prior social connection, and their interaction on correction engagement for lowest 3 bins of user partisan extremity ($N=1,115$).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.245	0.013	19.05	<0.001***
Social condition	0.055	0.026	2.14	0.033*
Shared partisanship	0.007	0.013	0.54	0.588
Social condition X Shared partisanship	-0.018	0.026	-0.71	0.477

Supplementary Table 70. Effect of shared partisanship, prior social connection, and their interaction on correction engagement for highest bin of user partisan extremity ($N=335$).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.312	0.025	12.34	<0.001***
Social condition	0.046	0.051	0.91	0.360
Shared partisanship	0.016	0.025	0.63	0.530
Social condition X Shared partisanship	0.118	0.051	2.33	0.020*

Supplementary Table 71. Effect of shared partisanship, prior social connection, and user partisan extremity dummy (0=lower 3 partisan extremity bins, 1=greatest partisan extremity bin) on engagement ($N=1,456$).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.245	0.013	18.705	<0.001***
Social condition	0.055	0.026	2.094	0.036*
Shared partisanship	0.007	0.013	0.532	0.595
User partisan extremity	0.067	0.027	2.461	0.014*
Social condition × Shared partisanship	-0.018	0.026	-0.698	0.485
Social condition × User partisan extremity	-0.010	0.054	-0.187	0.852
Shared partisanship × User partisan extremity	0.009	0.027	0.329	0.742
Social Condition × Shared partisanship × User partisan extremity	0.136	0.054	2.501	0.013*

Supplementary Table 72. Effect of shared partisanship, prior social connection, and user partisan extremity dummy (0=lower 3 partisan extremity bins, 1=greatest partisan extremity bin) on engagement for all users ($N=1,586$).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.238	0.011	22.378	<0.001***
Social condition	0.028	0.021	1.327	0.185
Shared partisanship	0.017	0.011	1.593	0.111
User partisan extremity	0.027	0.011	2.494	0.013*

Social condition × Shared partisanship	0.028	0.021	1.318	0.188
Social condition × User partisan extremity	-0.007	0.021	-0.315	0.753
Shared partisanship × User partisan extremity	0.007	0.011	0.690	0.490
Social Condition × Shared partisanship × User partisan extremity	0.064	0.021	2.994	0.003**

Supplementary Figure 7. Effect of prior social connection and shared partisanship on correction engagement for all users, by user partisan extremity (binned: low extremity = lowest 3 bins, high extremity = highest bin). Error bars reflect 95% confidence intervals.

4.3. Participant Partisan Extremity in Survey Experiment 1

Supplementary Table 73. Effect of shared partisanship, prior social connection, and participant partisan extremity on engagement (N=808).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.553	0.017	31.821	<0.001***
Social condition	0.089	0.035	2.559	0.011*
Shared partisanship	0.027	0.017	1.545	0.123
User partisan extremity	0.048	0.017	2.733	0.006**
Social condition × Shared partisanship	0.048	0.035	1.389	0.165
Social condition × User partisan extremity	0.001	0.035	0.30	0.976
Shared partisanship × User partisan extremity	0.015	0.017	0.854	0.394

Social Condition × Shared partisanship × User partisan extremity	0.017	0.035	0.476	0.634
--	-------	-------	-------	-------

Supplementary Table 74. Effect of shared partisanship, prior social connection, and their interaction on correction engagement for lowest 3 bins of participant partisan extremity (N=586).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.549	0.021	26.806	<0.001***
Social condition	0.101	0.041	2.460	0.014*
Shared partisanship	0.023	0.021	1.109	0.268
Social condition X Shared partisanship	-0.013	0.041	-0.315	0.753

Supplementary Table 75. Effect of shared partisanship, prior social connection, and their interaction on correction engagement for highest bin of user partisan extremity (N=222).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.564	0.033	17.200	<0.001***
Social condition	0.046	0.066	0.703	0.483
Shared partisanship	0.042	0.033	1.283	0.201
Social condition X Shared partisanship	0.201	0.066	3.062	0.002**

Supplementary Table 76. Effect of shared partisanship, prior social connection, and user partisan extremity dummy (0=lower 3 partisan extremity bins, 1=greatest partisan extremity bin) on engagement (N=808).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.549	0.020	26.924	<0.001***
Social condition	0.101	0.041	2.471	0.014*
Shared partisanship	0.023	0.020	1.114	0.267
User partisan extremity	0.014	0.039	0.370	0.711
Social condition × Shared partisanship	-0.013	0.041	-0.316	0.752
Social condition × User partisan extremity	-0.055	0.078	-0.703	0.483
Shared partisanship × User partisan extremity	0.019	0.039	0.496	0.620
Social Condition × Shared partisanship × User partisan extremity	0.214	0.078	2.743	0.006*

5. Survey Experiment S1a

5.1. Pre-registration.

Following our Twitter field experiment and prior to conducting survey experiment 1, we conducted an initial follow-up survey experiment (S1a), reported here ($N=1,290$, $M_{\text{age}} = 34.99$, 420 female). This survey experiment was pre-registered (https://aspredicted.org/blind.php?x=OMI_DAZ) – full analyses and results are presented here as pre-registered. In this survey experiment, participants were first given a series of up to 24 possible true and false headlines (12 each). Upon sharing a false headline, participants advanced onto the main treatment part of the survey ($N=948$). Participants then received a fictitious response to their shared article (baseline, social x co-partisan, counter-partisan; concordance determined from corrector partisanship and Clinton vs. Trump binary choice item) and indicated whether they would reply. If participants indicated they would reply, they were then asked to write their free-response reply. This written response was then coded as either ‘belief updating’ or ‘resisting information’ by Amazon Mechanical Turk workers in a separate study ($N=101$ raters, each reply rated 10.56 times on average, Kendall’s $W=0.186$, $ICC(1,k)=0.570$).

5.2. Descriptive Statistics.

Supplementary Table 77. Descriptive of engagements, belief updating responses, and resisting information responses by conditions ($N=948$) Belief updating and resisting information responses do not always equal total number of engagements, as some participants indicated they would engage, but did not write a free-response reply – as pre-registered, engagement was scored only using the binary choice reply item.

Experimental condition	<i>N</i>	Number of engagements	Number of positive responses	Number of negative responses
Baseline counter-partisan	248	116	61	40
Social counter-partisan	239	131	76	44
Baseline co-partisan	231	117	74	42
Social co-partisan	230	142	99	42

5.3. Main Analyses.

Supplementary Table 78. Effect of prior social connection and shared partisanship on engagement.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.537	0.016	33.216	<0.001***
Social condition	0.098	0.032	3.035	0.002**
Shared partisanship	0.050	0.032	1.533	0.126
Social condition X Shared partisanship	0.025	0.065	0.393	0.694

Supplementary Table 79. Effect of prior social connection and shared partisanship on belief updating.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.339	0.016	21.797	<0.001***

Social condition	0.093	0.031	2.983	0.003**
Shared partisanship	0.077	0.031	2.467	0.014*
Social condition X Shared partisanship	0.036	0.062	0.575	0.566

Supplementary Table 80. Effect of prior social connection and shared partisanship on resisting information.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.183	0.013	14.321	<0.001***
Social condition	0.012	0.026	0.467	0.641
Shared partisanship	-0.001	0.026	-0.038	0.970
Social condition X Shared partisanship	-0.022	0.051	-0.436	0.663

Supplementary Table 81. Significance of difference between coefficients of belief updating and resisting information.

	<i>p</i>
(Intercept)	<0.001***
Social condition	<0.001***
Shared partisanship	<0.001***
Social condition X Shared partisanship	0.099

Supplementary Table 82. Effect of prior social connection and shared partisanship on perceived positive corrector motive (1=corrector cares about truth; 0=corrector was politically motivated or trolling)..

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.423	0.016	26.272	<0.001***
Social condition	0.020	0.033	0.623	0.533
Shared partisanship	0.084	0.033	2.574	0.010*
Social condition X Shared partisanship	0.059	0.059	0.901	0.368

5.4. Pre-registered Secondary Analyses.

Supplementary Table 83. Effect of prior social connection and shared partisanship on replying to argue with the corrector.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.085	0.009	9.362	<0.001***
Social condition	-0.008	0.018	-0.449	0.654
Shared partisanship	-0.014	0.018	-0.773	0.439
Social condition X Shared partisanship	0.034	0.036	0.944	0.345

Supplementary Table 84. Effect of prior social connection and shared partisanship on replying to discuss more with the corrector.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.179	0.012	14.465	<0.001***

Social condition	0.052	0.025	2.091	0.037*
Shared partisanship	0.058	0.025	2.336	0.020*
Social condition X Shared partisanship	0.002	0.050	0.045	0.964

Supplementary Table 85. Effect of prior social connection and shared partisanship on replying to thank the corrector.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.248	0.014	17.752	<0.001***
Social condition	0.063	0.028	2.271	0.023*
Shared partisanship	0.055	0.028	2.972	0.049*
Social condition X Shared partisanship	-0.029	0.056	-0.520	0.603

Supplementary Table 86. Effect of prior social connection and shared partisanship on not replying because participant knows they were correct.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.087	0.009	9.488	<0.001***
Social condition	0.004	0.018	0.222	0.825
Shared partisanship	-0.027	0.018	-1.465	0.143
Social condition X Shared partisanship	-0.008	0.037	-0.204	0.838

Supplementary Table 87. Effect of prior social connection and shared partisanship on not replying because participant knows corrector was correct.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.050	0.007	7.029	<0.001***
Social condition	-0.010	0.014	-0.712	0.476
Shared partisanship	0.000	0.014	0.034	0.973
Social condition X Shared partisanship	-0.040	0.028	-1.423	0.155

Supplementary Table 88. Effect of prior social connection and shared partisanship on not replying to avoid confrontation.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.201	0.013	15.402	<0.001***
Social condition	-0.027	0.026	-1.022	0.307
Shared partisanship	-0.029	0.026	-1.098	0.272
Social condition X Shared partisanship	0.038	0.052	0.720	0.472

Supplementary Table 89. Effect of prior social connection and shared partisanship on not replying because corrector was a stranger.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.122	0.011	11.512	<0.001***
Social condition	-0.065	0.021	-3.073	0.002**

Shared partisanship	0.011	0.021	0.536	0.592
Social condition X Shared partisanship	-0.016	0.042	-0.375	0.708

Supplementary Table 90. Effect of prior social connection and shared partisanship on self-reported belief updating.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.540	0.016	33.642	<0.001***
Social condition	0.025	0.032	0.793	0.428
Shared partisanship	0.130	0.032	4.049	<0.001***
Social condition X Shared partisanship	-0.124	0.064	-1.928	0.054

Supplementary Table 91. Effect of prior social connection and shared partisanship on self-reported resisting information.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.424	0.016	26.424	<0.001***
Social condition	-0.012	0.032	-0.387	0.699
Shared partisanship	-0.058	0.032	-1.820	0.069
Social condition X Shared partisanship	0.098	0.064	1.522	0.128

Supplementary Table 92. Effect of prior social connection, shared partisanship, and participant partisanship (z-scored; -0.5=Democrat, 0=Independent, 0.5=Republican) on engagement.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.524	0.017	31.719	<0.001***
Social condition	0.103	0.033	3.120	0.002**
Shared partisanship	0.076	0.033	2.293	0.022*
Participant partisanship	0.020	0.017	1.237	0.216
Social condition × Shared partisanship	-0.003	0.066	-0.047	0.963
Social condition × Participant partisanship	0.022	0.033	0.669	0.503
Shared partisanship × Participant partisanship	0.043	0.033	1.296	0.195
Social condition × Shared partisanship × Participant partisanship	-0.034	0.066	-0.517	0.605

Supplementary Table 93. Effect of prior social connection, shared partisanship, and participant partisanship (z-scored; -0.5=Democrat, 0=Independent, 0.5=Republican) on belief updating.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.342	0.016	21.686	<0.001***
Social condition	0.087	0.032	2.759	0.006**
Shared partisanship	0.073	0.032	2.322	0.020*
Participant partisanship	0.021	0.016	1.306	0.192

Social condition × Shared partisanship	0.038	0.063	-0.606	0.545
Social condition × Participant partisanship	0.006	0.032	0.178	0.859
Shared partisanship × Participant partisanship	0.013	0.032	0.399	0.690
Social condition × Shared partisanship × Participant partisanship	-0.068	0.063	-1.084	0.279

Supplementary Table 94. Effect of prior social connection, shared partisanship, and participant partisanship (z-scored; -0.5=Democrat, 0=Independent, 0.5=Republican) on resisting information.

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.179	0.013	13.939	<0.001***
Social condition	0.014	0.026	0.557	0.578
Shared partisanship	0.005	0.026	0.188	0.851
Participant partisanship	0.021	0.013	0.066	0.948
Social condition × Shared partisanship	0.001	0.051	-0.727	0.467
Social condition × Participant partisanship	-0.037	0.026	0.610	0.542
Shared partisanship × Participant partisanship	0.033	0.026	1.300	0.194
Social condition × Shared partisanship × Participant partisanship	0.036	0.051	0.696	0.487

Supplementary Table 95. Effect of prior social connection and shared partisanship on engagement (where engagement = 0 if participant said they would reply, but did not actually write a free-response reply).

	<i>b</i>	<i>Std</i>	<i>t</i>	<i>p</i>
(Intercept)	0.508	0.016	31.482	<0.001***
Social condition	0.099	0.032	3.064	0.002**
Shared partisanship	0.105	0.032	3.260	0.001**
Social condition X Shared partisanship	0.011	0.065	0.175	0.861

References.

- 1 Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science* **26**, 1531-1542 (2015).
- 2 Jordan, J., Yoeli, E. & Rand, D. Don't get it or don't spread it? Comparing self-interested versus prosocially framed COVID-19 prevention messaging. *PsyArXiv* **10** (2020).