scientific reports

Check for updates

OPEN

REGISTERED REPORT

A path forward on online misinformation mitigation based on current user behavior

Catherine King✉, Samantha C. Phillips & Kathleen M. Carley

Social media misinformation has become a serious societal problem, and recent research has focused on developing effective ways to counter its harmful impacts. This work investigates user-level countermeasures, or how individuals who see the misinformation respond to it directly, possibly to help stop its spread in their online communities. Using a registered report design, we conducted an online survey of 1010 American social media users who use social media at least once weekly. Participants were asked how they respond and think others *should* respond to misinformation they unintentionally post or see posted by others, and how their responses differ depending on their relationship with the person who posted that misinformation. Overall, the results revealed a difference between respondents' beliefs and actions: participants reported expecting others to exert more effort when responding to misinformation than the level of effort they themselves reported. Additionally, on average, participants were more likely to say they intervened when misinformation was posted by someone close to them rather than by an acquaintance or a stranger. Understanding current behavioral patterns and public opinion can inform efforts to elicit public participation in countering misinformation and increase the effectiveness of platform-level countermeasures.

Protocol registration: The stage 1 protocol for this Registered Report was accepted in principle on March 13th, 2024. The protocol, as accepted by the journal, can be found at: https://figshare.com/s/683b1e7c2f2bad96f604.

Since the foreign interference observed during the 2016 U.S. presidential election, there has been an increased research focus on the spread, impact, and mitigation of misinformation online[1–5]. Broadly, misinformation refers to information that is false, inaccurate, and/or misleading[6]. This includes falsehoods or deceptive content deliberately spread and those shared by accident (i.e., disinformation and misinformation[5,7]), as well as various types (e.g., trolling, fake news, rumors[6–8]). While misinformation disproportionately affects certain vulnerable communities[9–11], anyone is susceptible to some degree to share and believe in it.

Social media is thought to aid the dissemination of misinformation[2,4], and researchers are growing more concerned about how social media may be contributing to political polarization and distrust in institutions and the media. As society becomes increasingly misinformed, hostility towards partisan opposition can undermine civil discourse[5]. Differences in misinformation exposure can result in entirely different perceptions of reality and behavioral effects[12,13]. Therefore, the impact of misinformation on social media on political polarization and institutional trust is of profound concern.

Countering misinformation is a challenging problem partly because platforms often only allow researchers to have limited or selective access to social media data, especially data that could be used to evaluate the effectiveness of various countermeasures[3,14]. According to a review of 223 countermeasures studies since 1972 by Courchesne et al. (2021), there has been a disproportionate amount of research on the effects of fact-checking[15,16], debunking[17,18], and prebunking[19,20]. These countermeasures are likely overstudied because experiments on these interventions can be conducted without access to platform data[3]. Yet many countermeasures, including those that could target creators of disinformation like reporting, have been understudied because platforms typically do not share the relevant data with researchers[3].

Recent research has begun investigating the relationship between seeing misinformation countermeasures online and public perception of those countermeasures[21]. However, the experience of observing or posting misinformation differs from the experience of observing or conducting a countermeasure. Viewers of misinformation on social media can either directly confront the authors of misinformation by offering social corrections or indirectly counter the misinformation by, for example, reporting the misinformation. Studying individual behavior in response to seeing misinformation is critical because previous research has shown that

Software and Societal Systems Department, Carnegie Mellon University, Pittsburgh, USA. ✉email: cking2@cs.cmu.edu

debunking myths is more effective when it comes from a trusted source, like a friend or family member[18,22]. This suggests that individuals responding directly to misinformation from users in their network can help slow or even stop the spread of misinformation.

This work investigates if these social corrections happen and if they depend on the nature of the relationship between the misinformation poster and the observer. Given the scale in which social media companies must detect and respond to misinformation[23], users can play a crucial role in limiting the spread in real time. Indeed, social media companies such as X (formerly known as Twitter) have piloted programs like Community Notes where users can add corrections and/or context to tweets they deem misleading[24]. Additionally, previous research has shown that most people do not intend to spread misinformation[25,26]. Instead of consciously sharing misinformation, cognitive and socio-affective mechanisms (e.g., intuitive thinking, identity motives) facilitate sharing and even belief in misinformation in some cases[27,28]. If this is the case, nudging social media users to focus on accuracy goals could help limit the unintentional spread of misinformation[29].

In addition, we explore what people believe *others* should do when they see misinformation or post misinformation themselves. This provides researchers and policymakers a sense of what social media users want the norm response to be, which is essential for public outreach about crowdsourced misinformation mitigation. We also examine how expectations vary from reported actions to understand the extent to which people currently feel empowered to respond to misinformation regardless of (their own) situational constraints. Participants may want others to respond to misinformation with higher effort actions than they do themselves. This act of hypocrisy, failing to practice what one preaches[30], could be leveraged to induce prosocial behavior changes (e.g., directly addressing content they believe contains misinformation)[31–33]. Making the discrepancy between behavior and advocated norms for behavior salient can activate threats to self-integrity, driving behavior in line with advocated norms to minimize the dissonance[32].

This paper surveyed 1,010 United States residents who use social media at least weekly. This survey covered the social media platforms where participants encounter misinformation, if they have posted misinformation (intentionally or unintentionally), their response to seeing or posting misinformation, and their opinions on how they think others should respond to seeing or posting misinformation. Participants report their response and the response they expect others to do when seeing misinformation posted by someone else for three levels of closeness to the sender of misinformation: close (e.g., a close friend or family member), somewhat close (e.g., acquaintances, colleagues, friends, extended family), or not close (e.g., someone you do not know offline). We included seven ways to counter misinformation posted by others, four ways to respond to misinformation posted by oneself, and an option to engage in no action. These responses capture a broad range of actions that involve directly and indirectly interacting with the content containing misinformation. Our study involves the following research questions, also summarized in Table 1.

RQ1 How do people respond and think others should respond when they see misinformation? Do response(s) change based on how close the participant is to the poster of misinformation?

RQ2 How do people respond and think others should respond when they realize they have posted misinformation?

RQ3 How do people's responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?

RQ4 How do beliefs about responses to misinformation differ based on various demographic factors?

Social media users may refrain from directly responding to (recognized) misinformation due to a myriad of constraints, such as concerns over damaging interpersonal relationships or their own credibility[34]. In addition, verifying and correcting misinformative claims is a time-consuming, effortful process in practice[35]. Users may also feel helpless to counter misinformation given the vast amount available online[34]. We generally expect people will incorporate these constraints more when reporting their own response to misinformation than when describing expectations for others due to cognitive distortions like fundamental attribution error[36,37]. While users can account for the factors that drive their decision-making about how to respond to misinformation online, it is significantly more challenging to incorporate the hypothetical situational constraints of others. Moreover, asserting that others should respond with high levels of effort can uphold feelings of morality even if participants do not want to engage in the moral behavior (i.e., responding actively to misinformation online) for whatever reason[38,39]. Therefore, we hypothesize that people believe others should expend more effort to respond to misinformation online than they actually do (**H1.1**). Additionally, we expect people to think that others should spend more effort to respond to the misinformation they posted themselves than what they actually do (**H2.1**).

Previous work shows users are more likely to correct a close contact because it is perceived as more worthwhile[34]. If users are going to take the time to engage with misinformative content directly, they want to feel like it will have an impact. If they have a personal relationship with the sender of misinformation, they have more information about the expected effectiveness of their correction. Furthermore, people may be especially concerned about close contacts believing in misinformation due to the potential negative consequences. We expect this will translate to expectations for others as well. Hence, we hypothesize that people respond with more effort when the sender of misinformation is a close contact than a somewhat close contact and a somewhat close contact than a not close contact (**H1.2**). Additionally, we expect people to want others to respond in the same way as H1.2 (**H1.3**).

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis plan	Interpretation given to different outcomes
RQ1: How do people respond and think others should respond when they see misinformation?	H1.1: People believe individuals should expend more effort to respond to misinformation online than they actually do	The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. The estimated number of participants who have seen misinformation per closeness level is ~ 192	• Calculate Measures 1a and 1b at each of the three closeness levels. Measure 1a: Determine the maximum level of effort individuals engage in when seeing misinformation posted by others. See the methods section and Table 2 for more details. Measure 1b: Determine the maximum level of effort individuals say other people should do • Run three one-sided Bayesian paired hypothesis tests (one at each closeness level) • The null hypothesis is that Measure 1b < = Measure 1a and the effect size [d = (Measure 1b – Measure 1a)/standard deviation] is < = 0. So, the null interval range is -Inf to 0. The alternate hypothesis is the effect size is > 0 • Calculate the highest density interval (HDI) for the effect size. Create a bar chart or another visualization that compares possible actions (See Table 2) in both scenarios	• The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true. For example, a Bayes factor of 10 indicates that the data are 10 times more likely to come from H1 than from H0 • The visualizations will show readers which actions were the most selected by survey participants
RQ1: How do people respond and think others should respond when they see misinformation?	H1.2: People respond with more effort when the sender of misinformation is a close contact than a somewhat close contact and a somewhat close contact than a not close contact	The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. The estimated sample size is n = 115	• Determine Measures 1a at each of the three closeness levels. See the methods section and Table 2 for more details • Run a one-sided Bayesian paired null hypothesis test comparing the effort level (Measure 1a) for close contacts and somewhat close contacts. Run a one-sided Bayesian paired null hypothesis test comparing the effort level (Measure 1a) for somewhat close contacts and not close contacts. For both tests, the null interval is -Inf to 0. The alternate hypothesis is that the effect size is greater than 0 • Calculate the HDI and create relevant visualizations	• The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true • Visualizations will help the reader interpret the results
RQ1: How do people respond and think others should respond when they see misinformation?	H1.3: People believe others should respond with more effort when the sender of misinformation is a close contact than a somewhat close contact and a somewhat close contact than a not close contact	The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. All participants will be included in this test, indicating a sample size of approx. n = 1000	• Determine Measures 1b at each of the three closeness levels. See the methods section and Table 2 for more details • Run a one-sided Bayesian paired null hypothesis test comparing the effort level (Measure 1b) for close contacts and somewhat close contacts. Run a one-sided Bayesian paired null hypothesis test comparing the effort level (Measure 1b) for somewhat close contacts and not close contacts. For both tests, the null interval is -Inf to 0. The alternate hypothesis is that the effect size is > 0 • Calculate the HDI and create relevant visualizations	• The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true • Visualizations will help the reader interpret results
RQ2: How do people behave after realizing they have posted misinformation	H2.1: People believe others should expend more effort to respond to misinformation online after realizing they posted misinformation than what they actually do	The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. The estimated number of participants who have posted misinformation accidentally is n = 320	• Measure 2a: Determine the maximum level of effort individuals engage after realizing they have posted misinformation. See methods section and Table 3 for more details. Measure 2b: Determine the maximum level of effort individuals say people should do when they realize they have posted misinformation • Run a one-sided Bayesian paired test with these two measures The null hypothesis is that the effect size is < = 0, so the null interval is -Inf to 0. The alternate hypothesis is the effect size is > 0 • Calculate the HDI and create relevant visualizations	• The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true • Visualizations will help the reader interpret results
RQ3: How do people's responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?	H3.1: People respond with a different level of effort when the sender of misinformation is someone else compared to themselves	The necessary sample size for a two-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 92. The estimated number of participants is 150	• Use Measures 1a and 2a (as described in H1.1, H2.1). See methods section and Tables 2 and 3 for more details • Run a two-sided paired Bayesian test with these two measures (null hypothesis is these two measures are equal and the effect size is effectively 0. Alternate hypothesis is that they differ). The null interval is – 0.2 to 0.2 (effect size is effectively 0). The alternate hypothesis is that the absolute value of the effect size is at least 0.2 • Calculate the HDI and create relevant visualizations	• The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true • If the 95% highest density interval falls entirely within the "region of practical equivalence" (ROPE) range, of – 0.2 to 0.2, then we can accept the null. If it falls entirely outside the range, then we can accept the alternate. Otherwise, using this method we will state inconclusive results • Visualizations will help the reader interpret results
RQ3: How do people's responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?	H3.2. People want others to respond with a different level of effort when the sender of misinformation is someone else compared to themselves	The necessary sample size for a two-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 92. All participants will be included in this test, indicating a sample size of approx. n = 1000	• Use Measures 1b and 2b (as described in H1.1, H2.1). See methods section and Tables 2 and 3 for more details • Run a two-sided paired Bayesian test with these two measures (null hypothesis is these two measures are equal, with an effect size of effectively 0. Alternate hypothesis is that they differ) The null interval is – 0.2 to 0.2. The alternate hypothesis is that the absolute value of the effect size is at least 0.2 • Calculate the HDI and create relevant visualizations	• The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true • If the 95% highest density interval falls entirely within the "region of practical equivalence" (ROPE) range, of – 0.2 to 0.2, then we can accept the null. If it falls entirely outside the range, then we can accept the alternate. Otherwise, using this method we will state inconclusive results • Visualizations will help the reader interpret results

Table 1. Design table.

When comparing responses to misinformation posted by others and posted by oneself, many individual and social factors may come into play, including wanting to preserve harmony and credibility or avoid embarrassment. Previous research from Singapore has shown that many young people avoid correcting misinformation posted by others to maintain their interpersonal relationships but do correct themselves to preserve their credibility despite possible embarrassment[40]. Other work suggests that many only correct others if they are close contacts or when it is an issue they care about[34]. Due to the conflicting literature in this area, we have created a non-directional hypothesis where we expect that people respond with a different level of effort when the sender of misinformation is someone else compared to themselves (**H3.1**). We also expect that people want others to respond with a different level of effort when the sender of misinformation is someone else compared to themselves (**H3.2**).

Finally, we investigate how behavior and beliefs about responses to misinformation on social media vary by partisanship and other demographic factors (RQ4). Extensive previous research has examined differences in misinformation susceptibility across age, gender, education level, income bracket, religious groups[11,41–44], and partisan groups[45,46], as well as the effectiveness of interventions across demographic groups[47]. Increasingly, researchers are studying how individual factors impact support for misinformation interventions, typically at a platform level[21,48–50]. In specific contexts, such as highly partisan environments, less susceptibility to misinformation is associated with more support for platform interventions. Left-leaning, Democratic individuals are both more supportive of platform interventions[21,51,52], and less likely to observe or spread misinformation online[45,46], than right-leaning or Republican individuals. In other cases, high susceptibility is linked to more support. Older adults are typically associated with higher susceptibility to sharing and believing misinformation[53,54]. Yet there is evidence that they support nudge interventions more[48].

Crucially, the type of misinformation seems to substantially affect susceptibility and views of countermeasures (e.g., older adults seem less susceptible to health-related misinformation than younger people[11]). Given the variability in susceptibility and support for platform-level interventions, we conducted exploratory analyses on how demographic attributes and partisanship affect support for and engagement with individual interventions. Understanding how specific populations, particularly those shown to be highly susceptible to misinformation, view and enact individual interventions informs effective public messaging about countering harmful content online.

Methods

Ethics information

The Institutional Review Board of Carnegie Mellon University approved this survey, numbered "STUDY2022_00000143." They approved this study as exempt from a full review because it is a survey that did not collect personally identifiable information. All methods were performed in accordance with the relevant guidelines and regulations. Informed consent was obtained from all participants. We expected the survey to take 15–18 min based on pilot tests. Participants were paid $3 each, which is equivalent to $10/hour if they took 18 min to complete the survey.

Pilot data

The survey was implemented in Qualtrics and was sent out to a small sample of Cloud Research Mechanical Turk participants to ensure questions were straightforward and the bot and duplicate detection worked. Twenty-two participants attempted the survey: 14 were excluded, most of them automatically by Cloud Research, for either being spam/bots, being a duplicate response, or failing to pass the screening questions (18 +, U.S. resident, use social media weekly). Participants excluded for these reasons were removed at the beginning of the survey and were not paid.

Participants were also asked if they had any comments. As a result, a few questions were removed to prevent the survey from being too long or reworded for succinctness and clarity. This document's hypotheses and research questions are based on the revised survey. While a pilot sample of eight responses is small, it helped improve the research design. It demonstrated that the questions were understandable and that this survey could effectively address the hypotheses and research questions. See the Supplementary Information in the Stage 1 Protocol for more detailed information on the pilot data.

Design

The survey was designed to answer this document's research questions and hypotheses. There are additional related questions on this survey that are not used in this study. Those questions are pre-registered elsewhere for a different study on countermeasures. The survey is included as a supplemental file.

Sampling plan

This section describes our sample characteristics, sample size determination, data exclusion criteria, all primary measures, and power analysis.

Participants

Our survey had 1,010 participants, and the data was collected between July and August 2024. This sample size was deemed appropriate because it provided sufficient power for our proposed hypotheses. Our survey was implemented using Qualtrics and administered through Cloud Research, an online recruiting platform[55], using Mechanical Turk survey participants. Only those respondents who are United States residents, adults, and use social media at least once a week were given the entire survey.

Response	Effort level
Ignore the post	No effort
Report the post	Low effort
Report the user	Low effort
Unfollow or unfriend the user	Low effort
Block the user	Low effort
Privately message the user	High effort
Comment a correction on the post	High effort
Create a separate post with the correct information	High effort

Table 2. Actions social media users can take when they see misinformation online.

Response	Effort level
Leave post as is	No effort
Delete the post	Low effort
Comment a correction on the post	High effort
Update the main post with a correction	High effort
Create a new post with the correct information	High effort

Table 3. Actions social media users can take when they realize they have posted misinformation online.

Procedure

Qualifying questions

We employed several methods to recruit relevant participants and maintain high data quality. Participants were adult U.S. residents who use social media weekly, and they also must have met the following criteria for inclusion:

1. Approved by Cloud Research
2. Had a higher than 95% approval rating on Mechanical Turk
3. Finished the survey
4. Not a bot (both Qualtrics and Cloud Research have bot detection)[56]
5. Not a duplicate response (Qualtrics flags likely duplicate responses)[56]

Data that failed even one of these criteria was excluded. A question at the end of the study asked if participants answered randomly at any point. This measure was for data quality purposes only and was not used to exclude data. 997 participants (99.7%) responded with "no", while 3 participants responded with "yes" (0.3%), and 10 skipped the question. The median time to complete the survey was 11.0 min, while the mean was 14.3 min.

Previous research has suggested that Mechanical Turks' data quality is high and that Turkers are more likely to pass attention-checking questions than other online panels[57]. A previous study also shows that the 95% approval rate cut-off can ensure high-quality data without using attention-checking questions[58]. Therefore, this survey did not have any attention-checking questions.

Behavioral questions

This section asked participants how they respond to seeing misinformation on social media platforms they use and how they react if they realize they have posted misinformation.

First, participants were asked if any of their social media contacts have ever posted something they believe to be misinformation, how often they saw it, and on which platforms they saw it. They were able to select among the top 11 most frequently used platforms in the United States as determined by Pew Research[59]. They also had the option to write in another platform that was not listed. Then, for all platforms they claimed to have seen misinformation on, they were asked how close they were to the people posting misinformation and how they responded. Possible responses are shown in Table 2. While these questions were broken up by which platform they saw the misinformation on, platform selection is not used for this study, and the data are summarized by closeness level. Platform analysis is saved for future work.

Next, participants were asked if they had ever intentionally or unintentionally posted misinformation. If they have unintentionally posted misinformation, they were asked on which platforms and then asked what they did on each platform once they realized they posted misinformation. Possible actions they could have taken are shown in Table 3.

Belief questions

This section asked participants how they thought people should respond when seeing misinformation. The questions were broken up by closeness: how should people respond to misinformation posted by a close contact? A somewhat close contact? A not close contact? Again, possible responses they were able to select are described

in Table 2. Finally, participants were asked what people should do if they realize they have posted misinformation (possible actions are described in Table 3).

Demographic questions

This section asked participants for various demographic characteristics. These were age, gender, race, ethnicity, education, household income, religion, political party affiliation, and general political leanings.

Other measures

Measure 1a) Effort Expended to Respond to Misinformation Posted by Others (Actual Behavior)

Measure 1b) Effort Expended to Respond to Misinformation Posted by Others (Opinion)

To test Hypotheses 1.1, 1.2, and 1.3, we created measures 1a) and 1b) to quantify the amount of effort put into responding to misinformation posted by others online. Table 2 shows a list of possible responses one could have when seeing misinformation on social media, generalized to apply to various social media platforms, and rated as *no effort, low effort*, or *high effort.* The only *no-effort* response is ignoring the post. Respondents could also respond with "I don't remember," in which case their effort level was not recorded, as it is unknown. A *low-effort* response means an action was taken, but there was no interaction with the content directly. A *high-effort* response indicates that the user likely took more time to respond and interacted with the content directly. The participants were able to select more than one of these actions. Pilot data values for Measures 1a) and 1b) for somewhat close contacts are in the Supplementary Information file in the Stage 1 Protocol.

Measure 2a) Effort Expended to Respond to Misinformation Posted by Oneself (Actual Behavior)

Measure 2b) Effort Expended to Respond to Misinformation Posted by Oneself (Opinion)

To test Hypothesis 2.1, we created measures 2a) and 2b) to quantify the effort one puts into correcting misinformation they posted online. Table 3 shows a list of possible actions someone could take. They are rated in the same way as the efforts described in Table 2: *no, low*, or *high effort.* Like in Table 2, the only *no-effort* response is leaving the post as is. Respondents could also respond with "I don't remember," in which case their effort level was not recorded, as it is unknown. Deleting the post is categorized as *low effort.* The remaining actions are classified as *high effort,* as they indicate the user took more time to respond and they placed effort into correcting their mistake. Pilot data values for Measures 2a) and 2b) are in the Supplementary Information file in the Stage 1 Protocol.

Anything labeled in Tables 2 or 3 as *no effort* received a score of 0, *low effort* received a score of 1, and *high effort* a score of 2. Anything without a label was labeled as NA. Participants were given the value of the highest effort level they engaged in per closeness level.

Bayes factor design analysis

We used a Bayesian approach to test our hypotheses. Unlike frequentist methods and *p*-values, the Bayes factor can show evidence in favor of either the null or the alternate, not just "reject" or "fail to reject"[60,61]. Additionally, unlike a traditional confidence interval that gives the range of values that would not be rejected at a specified *p*-value, the highest density interval includes, say, the 95% high probable values for the estimated parameter[61]. Finally, corrections are typically needed for multiple t-tests due to a concern over the possible detection of false positives, and these corrections can result in reduced power[62]. However, Bayesian tests typically do not need a correction for multiple tests, as the Bayesian prior places a relatively high probability on null effects[63].

This work used a Bayes Factor fixed-n design, where n is our sample size. Given our budget, our maximum sample size was determined to be approximately 1000 respondents. We used the methods described in the Schönbrodt and Wagenmakers (2018) design analysis paper[60] and the BFDA R package[64] to run a strength of evidence analysis. Like a traditional power analysis in a classical frequentist approach, a strength of evidence analysis can estimate the sample size needed so that a strong Bayes factor is found in a specific percentage of studies. For this analysis, the Bayes factor threshold was set to 10 and 1/10. Researchers consider a Bayes factor of 10 to be in the "strong" evidence range[65], with 10 meaning that the data is 10 times more likely to follow H1 over H0, and 1/10 indicating the reverse[60]. For this design analysis, evidence with a Bayes factor within that range is deemed "inconclusive." The higher the Bayes factor, the lower the probability of receiving misleading evidence (false positives or false negatives)[60].

We used the JZS Bayes factor, which assumes that the effect of H1 follows a central Cauchy distribution[60,66]. The JZS Bayes Factor was selected because it is the recommended default when little is known about the expected effect size[60,66]. The effect size is Cohen's *d*, the difference in means divided by the standard deviation. The width parameter of the Cauchy distribution was set to $\sqrt{2}/2$, which is the recommended value if expecting smaller effect sizes and is the default used in many software packages, including the R Package BayesFactor[67]. It corresponds to expecting a 50% probability of an effect size between −0.707 and 0.707 for a two-sided test and between 0 and 0.707 for a one-sided test. Pilot data for Hypothesis 1.1 was calculated to have an effect size of 0.80, which we believe is not different enough from the default parameter to warrant changing it, especially with such a small pilot data set.

The strength of evidence analysis was run considering being able to detect a possible effect size of 0.5, which is traditionally interpreted as a "medium" effect size[68]. We chose a medium effect size since we believe a small effect size for our hypotheses would not translate into much practical significance. We used this effect size and a Bayes factor threshold of 10 for the following hypotheses to determine if our sample size was high enough to have a reasonable probability of obtaining strong evidence for an alternate H1. We used the BFDA R package[64], which uses a Monte Carlo method to determine this. It generates 10,000 random samples under H1 with an expected effect size of 0.5 and computes the Bayes factors for those runs given the JZS prior. We ran another 10,000 random samples under H0 and calculated the Bayes factors. Then, we used the distribution of the Bayes

factors found to determine the sample size necessary to achieve a high probability (95%) of achieving a Bayes factor of 10.

We found that our planned sample size of approximately 1,000 participants was likely sufficient for each hypothesis to detect a medium effect size at a Bayes Factor threshold of 10. See Appendix 1 in the Supplementary Information or the Stage 1 protocol for the details supporting this power analysis for each hypothesis.

Statistical analysis

In this section, we detail all pre-registered analyses. We ran paired hypothesis tests, treating the ordinal data for effort level as interval data. The literature is divided on whether treating ordinal data as interval is recommended[69–71]. Given the large sample size, we expect the interpretation will likely not change if we analyze the data as ordinal rather than interval. However, as supplemental work, we conducted robustness checks by analyzing a Chi-square test of independence (treating the data as categorical) to improve the robustness of our results. We performed all statistical tests using the BayesFactor R package[67] and the most recent R and R Studio versions at the time of the analysis (R v4.4.1[72], RStudio v 2024.04.2 + 764[73]).

Hypothesis 1.1 To test H1.1, we ran a one-sided paired Bayesian null hypothesis test comparing the effort level of participants' actions when seeing misinformation (Measure 1a) with the effort level participants say others should do when seeing misinformation (Measure 1b) for each closeness level. To run this analysis, we took each user's maximum effort level, ranging from 0 to 2, for each closeness level. The effect size is equal to (the mean of Measure 1b—mean of Measure 1a)/standard deviation. The null hypothesis (H0) is that the effect size is $<=0$ (the null interval range is -Inf to 0). The alternate hypothesis (H1) is that there is a difference in the means with an effect size of greater than 0. The 95% highest density intervals were also calculated.

Hypothesis 2.1 We tested this hypothesis in a similar manner as Hypothesis 1.1. We ran a one-sided paired Bayesian null hypothesis test comparing the effort level of participants' actions when realizing they have posted misinformation (Measure 2a) with the effort level participants say others should do when they realize they have posted misinformation (Measure 2b) for each closeness level. To run this analysis, we took each user's maximum effort level, ranging from 0 to 2, on each closeness level. We again calculated the 95% highest density interval.

Hypothesis 1.2 and 1.3 Similarly, we ran one-sided Bayesian hypothesis tests for these hypotheses, again using a null interval range of -Inf to 0. The highest density intervals were calculated.

Hypothesis 3.1 and 3.2 For Hypothesis 3.1, we ran a two-sided paired Bayesian hypothesis test comparing the effort level of participants' actions when seeing others post misinformation (Measure 1a) with the effort level of participants' actions when they realized they posted misinformation themselves (Measure 2a). Similarly, for Hypothesis 3.2, we compared Measure 1b and Measure 2b. For both, the null interval range was -0.2 to 0.2. We added a buffer of 0.2 because we consider an effect size that small to be practically equivalent to no effect size. We additionally calculated the highest density interval and made appropriate visualizations for all the results.

Results

We surveyed 1010 active social media users in the United States. Almost all the participants said they had seen misinformation on at least one social media platform. Table 4 shows the number of participants who had seen misinformation or admitted to posting it unintentionally. These numbers indicate how many participants were qualified to answer behavioral questions about what they do after seeing or posting misinformation.

Registered analyses

Table 5 summarizes the registered analyses for each of the hypothesis tests. For detailed descriptions of the hypotheses and analyses, refer to Table 1. The "Sample Size" column lists the number of participants whose responses qualified for each paired hypothesis test. The "Interpretation" column follows the standard classification scheme for Bayes factors[65].

RQ1: How do people respond and think others should respond when they see misinformation on social media?

First, we summarize the interventions that participants report using. Figure 1 shows the total number of participants who responded at least once with each possible intervention from Table 2. We see that ignoring the misinformation was the most common response at every closeness level. Higher-effort actions (like privately messaging the user, commenting, or creating another post) received relatively more traction when responding to misinformation against close contacts compared with somewhat or not close contacts.

For **H1.1,** we found overwhelming evidence (Bayes Factor > 100) that participants believe individuals should expend more effort responding to misinformation on social media than those individuals report actually doing

Question	Yes	No
Have you ever seen misinformation posted or distributed on social media?	93.3% (942)	6.7% (68)
Have you ever posted or linked to something you later realized was misinformation?	25.7% (260)	74.3% (750)

Table 4. Misinformation exposure.

Hypothesis	Sample size	Null interval	Mean difference (S.D.)	Effect size	Bayes factor	95% HDI for effect size	Interpretation	McNemar Chi-Sq test
H1.1: Compare close contacts	148	(-Inf, 0)	0.47 (0.88)	0.53	> 100	[0.35, 0.70]	Extreme evidence for H1	$p = 1e\text{-}7$
			0.99 (1.93)	*0.51*	*> 100*	*[0.33, 0.67]*	*Interpretation remains the same*	
H1.1: Compare somewhat close contacts	370	(-Inf, 0)	0.65 (0.92)	0.71	> 100	[0.59, 0.82]	Extreme evidence for H1	$p < 2e\text{-}16$
			1.34 (1.94)	*0.58*	*> 100*	*[0.47, 0.69]*	*Interpretation remains the same*	
H1.1: Compare not close contacts	880	(-Inf, 0)	0.45 (0.87)	0.51	> 100	[0.44, 0.58]	Extreme evidence for H1	$p < 2e\text{-}16$
			0.95 (1.87)	*0.51*	*> 100*	*[0.44, 0.58]*	*Interpretation remains the same*	
H1.2: Compare close vs. somewhat close contacts	122	(-Inf, 0)	0.21 (0.84)	0.26	> 100	[0.07, 0.43]	Extreme evidence for H1	$p = 1e\text{-}4$
			0.16 (1.56)	*0.11*	*6.94*	*[− 0.07, 0.28]*	*Moderate evidence for H1*	
H1.2: Compare somewhat close vs. not close contacts	327	(-Inf, 0)	− 0.052 (0.88)	− 0.059	0.17	[− 0.17, 0.050]	Inconclusive. Mod. evidence for H0, but not at BF 1/10 threshold	$p = 2e\text{-}4$
			− 0.21 (1.52)	*− 0.14*	*< 1/100*	*[− 0.24, − 0.027]*	*Extreme evidence for H0*	
H1.3: Compare close vs. somewhat close contacts	1010	(-Inf, 0)	0.10 (0.58)	0.18	> 100	[0.12, 0.24]	Extreme evidence for H1	$p = 3e\text{-}8$
			0.16 (1.34)	*0.12*	*> 100*	*[0.057, 0.18]*	*Interpretation remains the same*	
H1.3: Compare somewhat close vs. not close contacts	1010	(-Inf, 0)	0.42 (0.81)	0.51	> 100	[0.45, 0.58]	Extreme evidence for H1	$p < 2e\text{-}16$
			0.34 (1.78)	*0.19*	*> 100*	*[0.13, 0.25]*	*Interpretation remains the same*	
H2.1	256	(-Inf, 0)	0.30 (0.55)	0.55	> 100	[0.41, 0.68]	Extreme evidence for H1	$p = 1e\text{-}13$
			1.28 (1.94)	*0.66*	*> 100*	*[0.52, 0.79]*	*Interpretation remains the same*	
H3.1: Compare oneself vs. close contacts	49	(− 0.2, 0.2)	0.27 (0.86)	0.31	0.594	[0.010, 0.57]	Inconclusive. Anecdotal evidence for H0, but not at BF 1/10 threshold	$p = 1e\text{-}3$
			0.37 (2.26)	*0.16*	*0.125*	*[− 0.12, 0.43]*	*Interpretation remains the same*	
H3.1: Compare oneself vs. somewhat close contacts	133	(− 0.2, 0.2)	0.41 (0.88)	0.46	79.5	[0.27, 0.63]	Very strong evidence for H1	$p = 9e\text{-}9$
			0.56 (2.01)	*0.28*	*0.805*	*[0.10, 0.44]*	*Inconclusive*	
H3.1: Compare oneself vs. not close contacts	229	(− 0.2, 0.2)	0.52 (0.85)	0.61	> 100	[0.46, 0.74]	Extreme evidence for H1	$p < 2e\text{-}16$
			0.67 (1.80)	*0.37*	*31.7*	*[0.23, 0.50]*	*Very strong evidence for H1*	
H3.1: Compare oneself vs. max effort over all closeness levels	244	(− 0.2, 0.2)	0.31 (0.85)	0.36	26.8	[0.23, 0.49]	Strong evidence for H1	$p = 3e\text{-}13$
			0.30 (1.86)	*0.16*	*0.07*	*[0.031, 0.28]*	*Strong evidence for H0*	
H3.2: Compare oneself vs. close contacts	1010	(− 0.2, 0.2)	0.10 (0.75)	0.13	< 1/100	[0.071, 0.20]	Extreme evidence for H0	$p < 2e\text{-}16$
			0.50 (2.15)	*0.23*	*0.977*	*[0.17, 0.29]*	*Inconclusive*	
H3.2: Compare oneself vs. somewhat close contacts	1010	(− 0.2, 0.2)	0.20 (0.78)	0.26	6.92	[0.20, 0.32]	Inconclusive. Mod. evidence for H1 but not at BF 10 threshold	$p < 2e\text{-}16$
			0.76 (2.14)	*0.35*	*> 100*	*[0.29, 0.42]*	*Extreme evidence for H1*	
H3.2: Compare oneself vs. not close contacts	1010	(− 0.2, 0.2)	0.62 (0.84)	0.74	> 100	[0.67, 0.81]	Extreme evidence for H1	$p < 2e\text{-}16$
			1.71 (2.12)	*0.81*	*> 100*	*[0.74, 0.88]*	*Interpretation remains the same*	
H3.2: Compare oneself vs. max effort over all closeness levels	1010	(− 0.2, 0.2)	0.031 (0.68)	0.045	< 1/100	[− 0.016, 0.11]	Extreme evidence for H0	$p < 2e\text{-}16$
			0.20 (2.13)	*0.093*	*< 1/100*	*[0.031, 0.15]*	*Interpretation remains the same*	

Table 5. Pre-registered Bayesian paired hypothesis test results and interpretation. Values in italics are for the hypotheses being run with the effort level summed rather than the max as a robustness check. The *p*-value for the generalized McNemar's Chi-square Test of Independence is included as a second robustness check.

when they encounter misinformation. This extremely significant result held no matter how close the participant claimed to be to the poster of the misinformation (close contacts, somewhat close contacts, and not close contacts). The effect size was greater than 0.5 in all three closeness cases, indicating a moderate effect size. The unrestricted 95% highest posterior density interval for the effect sizes were [0.35–0.70], [0.59–0.82], and [0.44–0.58] for close, somewhat, and not close contacts, respectively. This result also held the same strength of evidence when the tests were run using a summed effort level rather than a maximum effort level. Figure 2A–C shows the distribution of the maximum effort level reported per closeness level.

For **H1.2**, we found that participants responded with more effort when the misinformation poster was a close contact vs. a somewhat close contact (BF > 100) but that there was little difference in responses for somewhat close contacts compared with not close contacts (BF inconclusive). However, for **H1.3,** we found that participants believe that more effort should be expended on close contacts compared with somewhat close contacts and somewhat close contacts compared with not close contacts (BF > 100). Despite the belief that more effort should be put into responding to misinformation posted by a somewhat close contact relative to a not close contact, participants treated their somewhat close contacts and not close contacts with similar effort levels in practice.

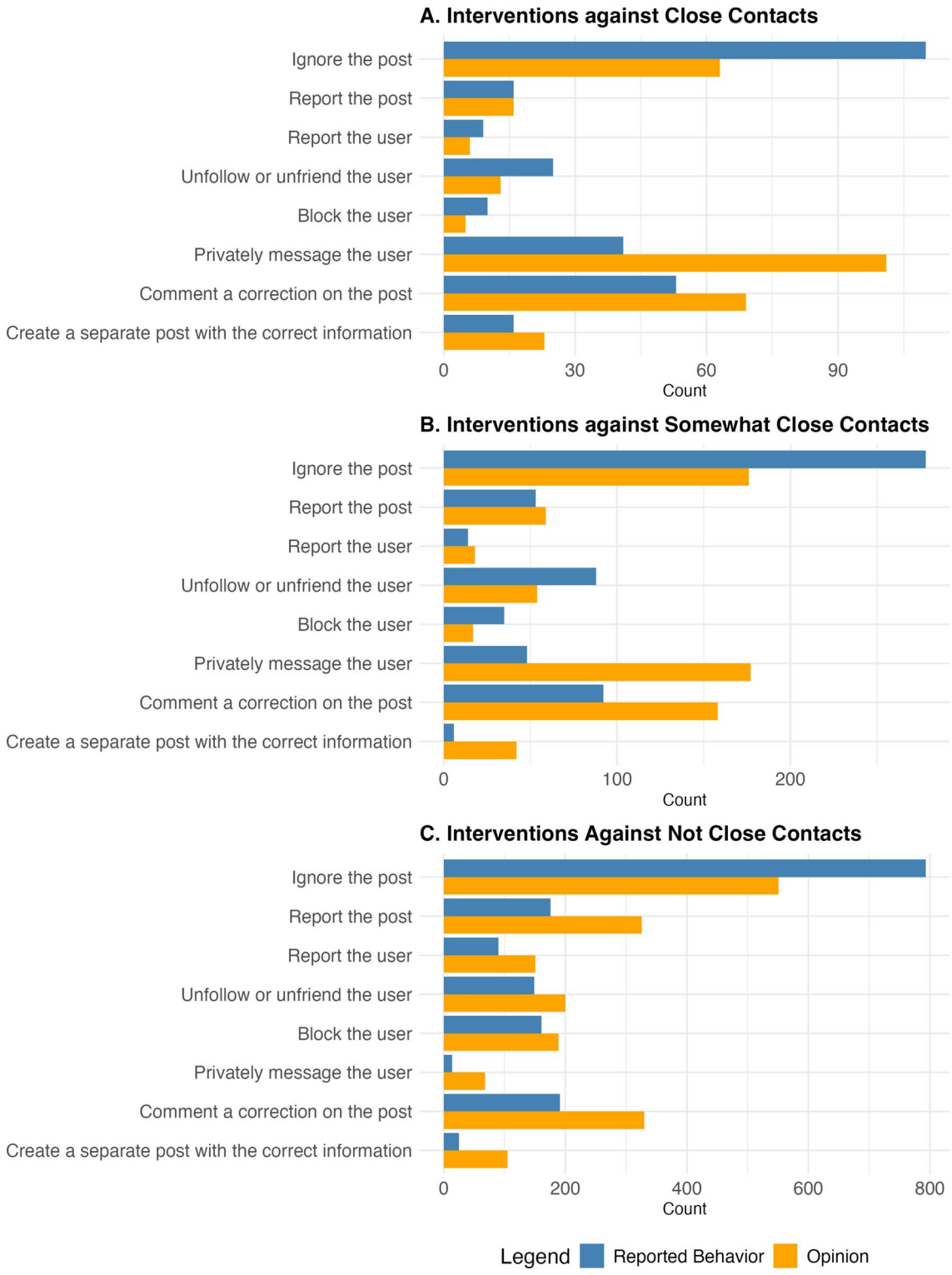

Fig. 1. Total number of participants who selected each intervention type from Table 2 at least once in their behavioral and opinion responses. For comparison purposes, only participants who had seen misinformation at that closeness level have their opinion counts included.

RQ2: How do people respond and think others should respond when they realize they have posted misinformation?

Over 25% of the participants (see Table 4) admitted to accidentally posting misinformation at least once. Figure 3 summarizes the interventions people claimed to have taken after realizing their mistake. The most frequently reported behavior was deleting the post, followed by updating the post with accurate information.

Fig. 2. Number of participants who reported expending a maximum of no, low, or high effort when seeing misinformation compared with the total number of those same participants who believe one should expend no, low, or high effort when seeing misinformation at each of the three closeness levels and against oneself.

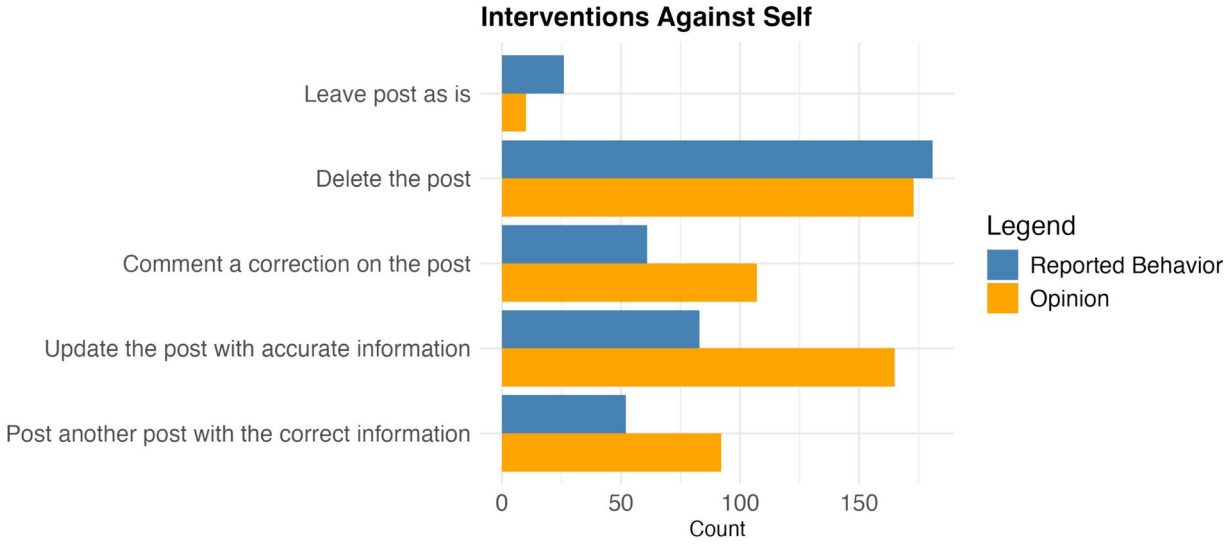

Fig. 3. Total number of participants who had posted misinformation and selected each intervention type from Table 3 at least once in their behavioral and opinion responses. For comparison purposes, only participants who admitted to posting misinformation had their opinion counts included.

We found extreme evidence (BF > 100) that people believe that they should expend more effort to respond to the misinformation they posted compared to what they actually do after realizing they have posted misinformation (**H2.1**). The effect size was 0.55, with the 95% HDI of the effect size being [0.41–0.68]. This result held with the same strength of evidence when the test was run using a summed effort level rather than a maximum effort level. See Fig. 2D for the distribution of maximum effort level after one has posted misinformation.

RQ3. How do people's responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?

We next investigated how participants' responses and beliefs differ when seeing misinformation versus posting it oneself. To answer this research question, we ran non-directional Bayesian hypothesis tests where we set the null interval to be between [− 0.2, 0.2]. For **H3.1,** we found strong evidence that participants respond with more effort when they post misinformation compared with when they see it posted by somewhat close (BF = 79.5) and not close contacts (BF > 100); however, inconclusive evidence that there was a difference in their responses when compared to close contacts.

For **H3.2,** we found extreme evidence (BF > 100) that participants *believe* that people should respond with more effort when they post misinformation compared to seeing it by not close contacts and inconclusive evidence (BF = 6.92) when comparing their misinformation posts to those posted by somewhat close contacts. Finally, we found very strong evidence (BF < 1/100) towards the null hypothesis that there is not a difference in the level of effort people believe one should use after posting misinformation oneself vs. seeing a close contact post it. These results indicate that participants believe the most effort should be afforded to counter misinformation posted by close contacts or themselves compared with countering misinformation posted by somewhat and not close contacts.

Robustness tests

We ran three robustness checks. First, we ran the registered hypothesis tests using summed effort values instead of maximum effort values. For H3.1 and H3.2, low-effort actions are excluded from this analysis because there are an unequal number of them described in Tables 2 and 3. In almost all cases, these tests yielded the same or a similar strength of evidence for the hypotheses. In the few instances where the results diverged, the registered test was inconclusive at the Bayes Factor threshold of 1/10 or 10, while the summed version of the test surpassed the threshold in the same direction (H1.2: somewhat vs. not close) or vice versa (H3.1: oneself vs. somewhat, H3.2 oneself vs. close). In only one instance did the interpretation completely differ: for H3.1 (oneself vs. max of all closeness), the registered test showed strong evidence for H1, whereas the summed test showed strong evidence for H0. Notably, the calculated effect size was positive in both cases. However, in the summed version of the hypothesis test, most of the 95% unrestricted highest density interval (HDI) lay below 0.2, placing it in the null interval.

Second, we used the generalized McNemar's Chi-square Test of Independence for categorical paired data to verify that the interpretation is similar if the data are analyzed categorically rather than as interval data. For every hypothesis, the chi-square test produced a p-value of < = 0.01. This outcome diverged from some of the pre-registered Bayesian analyses, which had found some inconclusive results or evidence pointing towards the null hypothesis for certain hypotheses. These discrepancies occurred in tests where the effect size was small, and the HDI overlapped with the null interval used in that test.

Finally, we ran a categorical analysis to analyze the association between the three variables of interest: maximum effort level used when countering, response type (reported behaviors vs. opinions), and closeness level. The three-way interaction term was not significant, indicating that closeness does not moderate the relationship between effort level and response type. Overall, the results were similar to those in our pre-registered analysis. Effort level interacts with both closeness level and response type, with higher counts of high-effort actions when contacts are closer or when people are asked about their opinions rather than their actual behavior. See Appendix 2 for additional details.

Exploratory analysis

RQ4. How do beliefs about responses to misinformation differ based on various demographic factors?

Finally, we examine individual differences in beliefs about how individuals should respond to misinformation. This exploratory analysis complements previous work that examines individual differences in support of misinformation countermeasures implemented by governments, social media companies, and other institutions[21]. For example, several previous studies have found that Democratic individuals are more supportive of platform interventions than Republicans[21,51,52], but does this translate into increased support for individual-level measures such as social corrections?

Figure 4 shows the maximum effort level participants believe one should exert when encountering misinformation posted by close contacts, somewhat close contacts, not close contacts, or oneself for six demographic variables: age, gender, race, education level, income level, and American political party. See Appendix 3 for details on the percentage of each demographic category that believes one should respond with no, low, or high effort when encountering misinformation posted by others or oneself. Appendix 3 also shows the detailed Chi-square test results for each demographic category and closeness level.

For political party affiliations, we find differences in belief in response efforts between partisan groups for close and not close contacts. Strong Republicans supported ignoring posts containing misinformation by close contacts more than any other group, although the absolute difference is < 10%, which may not have much practical significance. Furthermore, strong Democrats were more likely to support high or low-effort responses

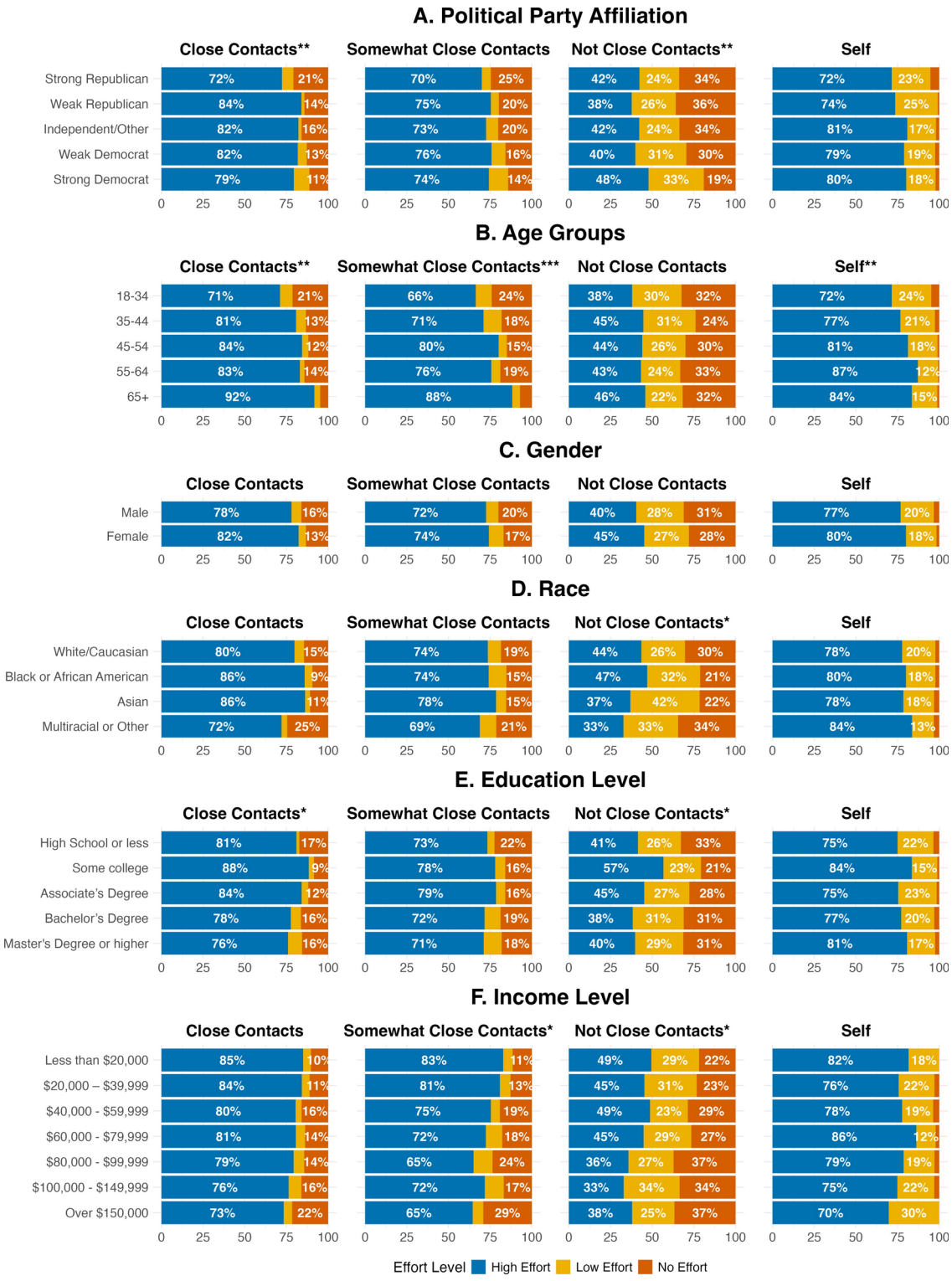

Fig. 4. Highest effort level participants said one should respond with when seeing misinformation posted by others or oneself broken up by party, age, gender, race, education, and income. Chi-sq tests: $p < 0.05$*, $p < 0.01$**, and $p < 0.001$***

to not close contacts more than any other group. Notably, we see that, except for not close contacts, at least 70% of respondents in all party affiliations said that one should respond with a high-effort action (such as commenting on a correction, updating the post, or messaging the poster).

For age, the chi-squared test shows statistically significant differences in responses among age groups when considering close contacts, somewhat close contacts, and oneself. In general, older participants were more likely

to believe one should exert a high level of effort when countering misinformation than younger participants. No significant differences were found between men and women. For racial groups, the only statistically significant difference found was for not close contacts, with Black and Asian Americans more likely to believe in responding with some effort than the other racial groups.

Finally, the percentage of American residents stating that one should use a high level of effort to counter misinformation drops as education or income level increases. The results from the chi-squared test show statistically significant differences in responses among various education groups regarding close contacts and not close contacts, and among various income groups regarding somewhat close contacts and not close contacts.

Discussion

In this registered study, we compared individuals' beliefs about ideal responses and actual responses to misinformation posted on social media by close contacts, somewhat close contacts, not close contacts, and themselves.

We found overwhelming evidence of hypocrisy in people's responses to misinformation, aligning with our hypotheses (H1.1, H2.1). Participants believe others should exert more effort to counter misinformation than they report doing themselves. This pattern holds across all closeness levels, including misinformation posted by oneself, and remains robust to multiple ways of measuring effort. Since there is already a widespread belief that individuals should combat misinformation, efforts to encourage social corrections do not have to convince people to support individual corrections. Instead, they can focus on normalizing these practices and providing strategies to overcome situational constraints (e.g., time and cognitive effort required, social pressures) preventing people from acting.

Furthermore, our results indicate that people not only expect others to exert more effort but also tend to invest more effort themselves when addressing misinformation posted by close contacts compared to those who are somewhat close or not close at all (H1.2, H1.3). This increased effort may stem from the impression that correcting a close contact is more likely to be effective due to their relationship, making the effort more worthwhile. Alternatively, people might feel a stronger sense of responsibility to correct a closer contact whose beliefs and behaviors could impact them offline. Additionally, the types of responses differ across closeness as well. For example, people are more likely to privately message a close contact than a less close one. Different approaches may feel more appropriate depending on the source of misinformation. Providing users with a range of options, including private or low-effort methods like reporting, may increase their likelihood of engaging in countering behavior.

When comparing responses to misinformation posted by oneself versus someone else of varying closeness (H3.1, H3.2), participants reported putting more effort into responding to misinformation they had posted than to misinformation posted by somewhat or not close contacts. Their beliefs about ideal responses also reflected this pattern. Interestingly, we also found strong evidence that individuals respond with similar levels of effort to misinformation they posted compared with misinformation posted by a close contact. This suggests a similar view of responsibility when the source of misinformation is oneself or a close contact.

Finally, our exploratory analysis revealed demographic differences in beliefs about countering misinformation. Strong Republicans were less likely to believe that high effort should be exerted when countering close contacts, whereas strong Democrats were more inclined to believe some level of effort should be used for not close contacts. This partially aligns with prior research indicating that strong Democrats stood out in their support for institutional countermeasures compared with other partisan groups[21]. We only find this difference holds for not close contacts. Individual-level interventions give people agency to respond to content they believe is misinformation, potentially mitigating distrust in institutional definitions of misinformation. Our findings suggest that this approach to addressing misinformation may be more palatable across the political spectrum.

We found that older Americans were more likely to believe that one should exert high effort to counter misinformation than their younger counterparts. This difference may reflect broader attitudes towards social media, as older individuals are more likely to perceive adverse effects associated with it[74] and, therefore, may be more motivated to address misinformation. Additionally, higher education and income levels were associated with a decreased belief that high effort should be exerted to counter misinformation. Interestingly, higher education and income are also associated with an increased concern and awareness of the negative impact of misinformation[25]. It may be that this concern does not necessarily translate into a belief in the effectiveness or necessity of individual countermeasures. Rather, these concerns may drive greater support for countermeasures on larger scales (e.g., government or platform), which is beyond the scope of this work but should be examined. Additionally, existing literature suggests that higher-income individuals are less generous overall[75,76], which may extend to efforts to counter misinformation. This preliminary exploratory work can inform future research and platform policies.

Implications

This work has several practical implications for promoting public participation in countering online misinformation in educational and technological contexts. First, our research demonstrates the widespread approval of social corrections online, indicating the social desirability of these behaviors. Prior work shows that highlighting the social desirability of reporting misinformation as an injunctive social norm can motivate reporting[77] and decrease the sharing of misinformation[78]. More broadly, there is strong evidence that social signals (e.g., engagement metrics and comments) influence responses to misinformation[79,80]. However, they can also encourage harmful behavior, such as sharing misinformation that aligns with one's pre-existing beliefs[81]. Therefore, platforms and organizations can successfully promote individual interventions by emphasizing their popularity among users, but they must be careful to avoid inadvertently empowering users with questionable motives.

Additionally, the observed disparity between reported behaviors and beliefs could be leveraged to encourage greater public participation. Research on hypocrisy suggests that one of the most effective strategies for driving behavioral changes is to have individuals publicly commit to pro-social actions, such as signing a pledge, and then be privately reminded of times they have failed to follow through[32]. Public call-outs are less effective, as they may prompt people to save face or rationalize their failures by reducing their support for the targeted behavior[32]. Social media platforms could encourage users, such as those who sign up to contribute to Community Notes programs, to publicly support social corrections or similar measures during educational sessions and regularly remind them of their commitment going forward.

Moreover, the result that people believe more should be done to respond to misinformation than they report doing themselves indicates that there may be barriers to employing social corrections or reporting features that could be mitigated by platform design, such as improving transparency, usability, and technical support of these features[82]. Platforms should educate users about their reporting systems when joining and provide periodic updates to keep them informed. Users are also more likely to use reporting features if they perceive them as effective. Therefore, sharing information about the outcomes of reports filed or community notes written can incentivize people to use these programs[82,83]. Furthermore, pop-up windows have been used on several platforms to ask users if they wish to share content they have not reviewed[84], and this method could be used in scenarios where users delete content. Instead of deleting potentially misleading posts, users could be encouraged to edit or update their posts with accurate information.

Recognizing that susceptibility to misinformation varies across demographic groups[11], educational efforts could tailor strategies to effectively reach different populations[85]. For example, older adults are particularly vulnerable to political misinformation, potentially due to lower digital literacy levels[53]. This age group also tends to support higher-effort responses to misinformation encountered online. Therefore, training efforts for older adults might prioritize digital media literacy over encouraging social corrections. People who are less vulnerable to misinformation, on the other hand, can be promptly educated about operationalizing corrections and leveraging specific platform affordances. Platforms could utilize their internal data to identify these users or implement reputation systems, like X's Community Notes program, where users earn "Rating Impact" scores based on the helpfulness of their contributions. (https://communitynotes.x.com/guide/en/contributing/writing-and-rating-impact) Overall, educational efforts should be designed to account for individual differences in both vulnerability to misinformation and perceived responsibility to counter it.

Lastly, there are a myriad of individual differences beyond demographics that influence vulnerability to misinformation and the likelihood to correct it that should be examined further in future work. For example, evidence suggests that those with a tendency toward analytical thinking are less susceptible to misinformation[50,86]. These are likely the same individuals capable of providing accurate and meaningful corrections, as some level of cognitive effort is necessary for higher-effort responses to misinformation (e.g., commenting on a correction). Platforms can encourage users to think critically by using accuracy prompts or similar measures[29,50]. In addition, platforms and other institutions can target educational resources towards those with a propensity to engage in critical thinking.

Limitations and future work

There are several limitations to this work. First, our sample is not demographically representative of all United States residents. We specifically focused on active social media users to better understand current user behavior on social media platforms. While this targeted sample provides relevant information to platforms about how their users act and what they believe, a more demographically representative survey could provide additional information about less active users who can also influence the spread of misinformation. Additionally, while we collected high-level demographic data, we did not investigate the role of more complex individual features, such as analytical reasoning or values, and how they may interact with one's propensity to intervene against misinformation. We leave this to future work.

Next, participants were asked to recall how they had responded to misinformation they had seen on social media in the past, which they may or may not have encountered recently. This could have led to memory or recall errors. Furthermore, we note the possibility of demand effects or other biases (e.g., social desirability) influencing survey responses. We took care to present the survey to participants as generally about misinformation online without including details that may reveal our expectations. Future work could consider using platform data or conducting a field experiment to observe how people respond to misinformation in real-world contexts. For example, platform data on reporting or social corrections could confirm whether people counter closer contacts more than less close contacts.

Additionally, the results linked to RQ3 may have limited generalizability due to the fundamental difference in the potential actions one can employ to correct others compared with correcting oneself. We attempted to enumerate commensurate responses to misinformation, such as commenting a correction on someone else's post or one's own post. However, the low-effort actions are, by nature, not equivalent actions (reporting someone else's post vs. deleting one's post). Additionally, there were more listed low-effort actions for responding to others than responding to oneself. We address this in one of our robustness checks, where we compared summed rather than maximum effort levels and excluded all low-effort responses. However, future studies should consider this inherent limitation when conducting this type of analysis.

Finally, we did not investigate possible differences among the platforms included in the study, which is an important future extension of this work. Future work should also consider expanding upon our exploratory work by investigating the behavior, not just the beliefs, of various demographic groups. Furthermore, it would be interesting to see if behavior or beliefs about how to engage on social media are related in any way to support for platform or government measures to counter misinformation.

Conclusion

This study makes an important contribution to the literature on individual-level interventions against misinformation. Our results indicate that facilitating individual responses to misinformation seen or accidentally posted on social media is a viable approach to reducing the spread of misinformation and even preventing belief in it. Using a large sample of active social media users in the US, we demonstrate the widespread belief that individuals should counter misinformation despite a tendency to not always act on this belief themselves. The nature of responses and the willingness to expend effort vary based on the user's relationship with the misinformation poster, highlighting opportunities to educate the broader population about different ways to take action depending on their perceived situational constraints. These insights inform efforts to encourage public participation in mitigating the impact of misinformation and suggest ways that platforms can enhance their countering tools to empower users to engage more actively in maintaining the integrity of their online information environment.

Data availability

The raw data and readme file are here: https://doi.org/10.1184/R1/27264786. The study materials are here: https://doi.org/10.1184/R1/27264813.

Code availability

The code can be found here: https://doi.org/10.1184/R1/27264780.

Received: 15 February 2023; Accepted: 3 March 2025
Published online: 19 March 2025

References

1. Seven ways misinformation spread during the 2016 election. *Knight Foundation* https://knightfoundation.org/articles/seven-ways-misinformation-spread-during-the-2016-election/ (2018).
2. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236 (2017).
3. Courchesne, L., Ilhardt, J. & Shapiro, J. N. Review of social science research on the impact of countermeasures against influence operations. *Harv. Kennedy Sch. Misinformation Rev.* **2**, (2021).
4. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374 (2019).
5. Tucker, J. et al. Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.3144139 (2018).
6. Wu, L., Morstatter, F., Carley, K. M. & Liu, H. Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explor. Newsl.* **21**, 80–90 (2019).
7. Wang, M., Rao, M. & Sun, Z. Typology, etiology, and fact-checking: A pathological study of top fake news in China. *J. Pract.* **16**, 1–19 (2020).
8. Zannettou, S., Sirivianos, M., Blackburn, J. & Kourtellis, N. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data Inf. Qual.* **11**, 1–37 (2019).
9. Jaiswal, J., LoSchiavo, C. & Perlman, D. C. Disinformation, misinformation and inequality-driven mistrust in the time of COVID-19: Lessons unlearned from AIDS denialism. *AIDS Behav.* **24**, 2776–2780 (2020).
10. Seo, H., Blomberg, M., Altschwager, D. & Vu, H. T. Vulnerable populations and misinformation: A mixed-methods approach to underserved older adults' online information assessment. *New Media Soc.* **23**, 2012–2033 (2021).
11. Nan, X., Wang, Y. & Thier, K. Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Soc. Sci. Med.* **314**, 115398 (2022).
12. Barua, Z., Barua, S., Aktar, S., Kabir, N. & Li, M. Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Prog. Disaster Sci.* **8**, 100119 (2020).
13. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* **7**, 74–101 (2023).
14. Assenmacher, D. et al. Benchmarking crisis in social media analytics: A solution for the data-sharing problem. *Soc. Sci. Comput. Rev.* **40**, 1496–1522 (2022).
15. 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 (ACM, New Orleans LA USA, 2022). https://doi.org/10.1145/3491102.3502040.
16. Bor, A., Osmundsen, M., Rasmussen, S. H. R., Bechmann, A. & Petersen, M. B. 'Fact-checking' videos reduce belief in misinformation and improve the quality of news shared on Twitter. Preprint at https://doi.org/10.31234/osf.io/a7huq (2020).
17. Ecker, U. K. H., Hogan, J. L. & Lewandowsky, S. Reminders and repetition of misinformation: Helping or hindering its retraction?. *J. Appl. Res. Mem. Cogn.* **6**, 185–192 (2017).
18. Lewandowsky, S., Cook, J. & Lombardi, D. Debunking Handbook 2020. Databrary https://doi.org/10.17910/B7.1182 (2020).
19. Lewandowsky, S. & van der Linden, S. Countering misinformation and fake news through inoculation and prebunking. *Eur. Rev. Soc. Psychol.* **0**, 1–38 (2021).
20. Roozenbeek, J., Linden, S. van der & Nygren, T. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harv. Kennedy Sch. Misinformation Rev.* **1**, (2020).
21. Saltz, E., Barari, S., Leibowicz, C. R. & Wardle, C. Misinformation interventions are common, divisive, and poorly understood. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-81 (2021).
22. Bode, L. & Vraga, E. K. See something, say something: Correction of global health misinformation on social media. *Health Commun.* **33**, 1131–1140 (2018).
23. Kaur, K. & Gupta, S. Towards dissemination, detection and combating misinformation on social media: A literature review. *J. Bus. Ind. Mark.* **38**, 1656–1674 (2022).
24. Wagner, K. Inside Twitter's plan to fact-check tweets. *Bloomberg.com* (2021).
25. Barthel, M., Mitchell, A. & Holcomb, J. Many Americans believe fake news is sowing confusion. *Pew Research Center's Journalism Project* https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/ (2016).
26. Watson, A. Sharing of made-up news on social networks in the U.S. 2020. *Statista* https://www.statista.com/statistics/657111/fake-news-sharing-online/ (2022).
27. Ecker, U. K. H. et al. The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* **1**, 13–29 (2022).

28. Chen, X., Sin, S.-C.J., Theng, Y.-L. & Lee, C. S. Why students share misinformation on social media: Motivation, gender, and study-level differences. *J. Acad. Librariansh.* **41**, 583–592 (2015).
29. Pennycook, G. et al. Shifting attention to accuracy can reduce misinformation online. *Nature* https://doi.org/10.1038/s41586-021-03344-2 (2021).
30. Brunsson, N. *The Organization of Hypocrisy: Talk, Decisions and Actions in Organizations* (Wiley, 1989).
31. Aronson, E., Fried, C. & Stone, J. Overcoming denial and increasing the intention to use condoms through the induction of hypocrisy. *Am. J. Public Health* **81**, 1636–1638 (1991).
32. Stone, J. & Fernandez, N. C. To practice what we preach: The use of hypocrisy and cognitive dissonance to motivate behavior change. *Soc. Personal. Psychol. Compass* **2**, 1024–1051 (2008).
33. Fointiat, V. 'I know what I have to do, but...' When hypocrisy leads to behavioral change. *Soc. Behav. Personal. Int. J.* **32**, 741–746 (2004).
34. Tandoc, E. C., Lim, D. & Ling, R. Diffusion of disinformation: How social media users respond to fake news and why. *Journalism* **21**, 381–398 (2020).
35. Southwell, B. G. et al. Misinformation as a misunderstood challenge to public health. *Am. J. Prev. Med.* **57**, 282–285 (2019).
36. Tetlock, P. E. Accountability: A social check on the fundamental attribution error. *Soc. Psychol. Q.* **48**, 227–236 (1985).
37. Langdridge, D. & Butt, T. The fundamental attribution error: A phenomenological critique. *Br. J. Soc. Psychol.* **43**, 357–369 (2004).
38. Lammers, J., Stapel, D. A. & Galinsky, A. D. Power increases hypocrisy: Moralizing in reasoning. *Immoral. Behav. Psychol. Sci.* **21**, 737–744 (2010).
39. Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C. & Wilson, A. D. In a very different voice: Unmasking moral hypocrisy. *J. Pers. Soc. Psychol.* **72**, 1335–1348 (1997).
40. Ng, S. W. T. Self- and social corrections on instant messaging platforms. (2023).
41. Bridgman, A. *et al.* The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harv. Kennedy Sch. Misinformation Rev.* **1**, (2020).
42. Carey, J. M., Chi, V., Flynn, D. J., Nyhan, B. & Zeitzoff, T. The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Sci. Adv.* **6**, eaaw7449 (2020).
43. Druckman, J. N. et al. The role of race, religion, and partisanship in misperceptions about COVID-19. *Group Process. Intergroup Relat.* **24**, 638–657 (2021).
44. Vijaykumar, S. et al. How shades of truth and age affect responses to COVID-19 (Mis)information: Randomized survey experiment among WhatsApp users in UK and Brazil. *Humanit. Soc. Sci. Commun.* **8**, 1–12 (2021).
45. González-Bailón, S. et al. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
46. Garrett, R. K. & Bond, R. M. Conservatives' susceptibility to political misperceptions. *Sci. Adv.* **7**, eabf1234 (2021).
47. Gwiaździński, P. *et al.* Psychological interventions countering misinformation in social media: A scoping review. *Front. Psychiatry* **13**, (2023).
48. Kirchner, J. & Reuter, C. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proc. ACM Hum. Comput. Interact.* **4**, 1–27 (2020).
49. Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychol. Sci. Public Interest* **21**, 103–156 (2020).
50. D'Errico, F., Cicirelli, P. G., Corbelli, G. & Paciello, M. Addressing racial misinformation at school: A psycho-social intervention aimed at reducing ethnic moral disengagement in adolescents. *Soc. Psychol. Educ.* https://doi.org/10.1007/s11218-023-09777-z (2023).
51. Mitchell, A. & Walker, M. More Americans now say government should take steps to restrict false information online than in 2018. *Pew Research Center* https://www.pewresearch.org/fact-tank/2021/08/18/more-americans-now-say-government-should-take-steps-to-restrict-false-information-online-than-in-2018/ (2021).
52. Kozyreva, A. et al. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proc. Natl. Acad. Sci.* **120**, e2210666120 (2023).
53. Brashier, N. M. & Schacter, D. L. Aging in an era of fake news. *Curr. Dir. Psychol. Sci.* **29**, 316–323 (2020).
54. Guess, A., Nyhan, B. & Reifler, J. Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. https://about.fb.com/wp-content/uploads/2018/01/fake-news-2016.pdf.
55. Hauser, D. J. et al. Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behav. Res. Methods* **55**, 3953–3964 (2023).
56. Fraud Detection. *Qualtrics Support* https://www.qualtrics.com/support/https://wordpressstaging.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/.
57. Hauser, D. J. & Schwarz, N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav. Res. Methods* **48**, 400–407 (2016).
58. Peer, E., Vosgerau, J. & Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* **46**, 1023–1031 (2014).
59. Auxier, B. & Anderson, M. Social Media Use in 2021. *Pew Research Center: Internet, Science & Tech* https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ (2021).
60. Schönbrodt, F. D. & Wagenmakers, E.-J. Bayes factor design analysis: Planning for compelling evidence. *Psychon. Bull. Rev.* **25**, 128–142 (2018).
61. Kruschke, J. K. & Liddell, T. M. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* **25**, 178–206 (2018).
62. Gelman, A., Hill, J. & Yajima, M. Why we (Usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* **5**, 189–211 (2012).
63. Neath, A. A., Flores, J. E. & Cavanaugh, J. E. Bayesian multiple comparisons and model selection. *WIREs Comput. Stat.* **10**, e1420 (2018).
64. Schönbrodt, F. D. & Stefan, A. M. BFDA: An R package for Bayes factor design analysis (version 0.5.0). (2019).
65. Lee, M. D. & Wagenmakers, E.-J. *Bayesian Cognitive Modeling: A Practical Course* (Cambridge University Press, Cambridge, 2013).
66. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).
67. Morey, R. D. *et al.* Package 'BayesFactor'. (2022).
68. Leppink, J., O'Sullivan, P. & Winston, K. Effect size – large, medium, and small. *Perspect. Med. Educ.* **5**, 347–349 (2016).
69. Robitzsch, A. Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Front. Educ.* **5**, (2020).
70. Wu, H. & Leung, S.-O. Can likert scales be treated as interval scales?—A simulation study. *J. Soc. Serv. Res.* **43**, 527–532 (2017).
71. Howcroft, D. M. & Rieser, V. What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (eds. Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.) 8932–8939 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021). https://doi.org/10.18653/v1/2021.emnlp-main.703.
72. R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/ (2024).
73. Posit team. *RStudio: Integrated Development Environment for R*. (2024).

74. Auxier, B. 64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today. *Pew Research Center* https://www.pewresearch.org/short-reads/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/ (2020).
75. Côté, S., House, J. & Willer, R. High economic inequality leads higher-income individuals to be less generous. *Proc. Natl. Acad. Sci.* **112**, 15838–15843 (2015).
76. Piff, P. K., Kraus, M. W., Côté, S., Cheng, B. H. & Keltner, D. Having less, giving more: The influence of social class on prosocial behavior. *J. Pers. Soc. Psychol.* **99**, 771–784 (2010).
77. Gimpel, H., Heger, S., Olenberger, C. & Utz, L. The effectiveness of social norms in fighting fake news on social media. *J. Manag. Inf. Syst.* **38**, 196–221 (2021).
78. Jones, C. M. et al. Impact of social reference cues on misinformation sharing on social media: Series of experimental studies. *J. Med. Internet Res.* **25**, e45583 (2023).
79. Avram, M., Micallef, N., Patil, S. & Menczer, F. Exposure to social engagement metrics increases vulnerability to misinformation. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-033 (2020).
80. Colliander, J. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Comput. Hum. Behav.* **97**, 202–215 (2019).
81. Lawson, M. A., Anand, S. & Kakkar, H. Tribalism and tribulations: The social costs of not sharing fake news. *J. Exp. Psychol. Gen.* **152**, 611–631 (2023).
82. Zhou, H., Lu, Y., Zhao, L., Wang, B. & Li, T. Effective reporting system to encourage users' reporting behavior in social media platforms: An empirical study based on structural empowerment theory. *Behav. Inf. Technol.* **43**, 3490–3509 (2024).
83. Wong, R. Y. M., Cheung, C. M. K., Xiao, B. & Thatcher, J. B. Standing up or standing by: Understanding Bystanders' proactive reporting responses to social media harassment. *Inf. Syst. Res.* **32**, 561–581 (2021).
84. Clark, M. Facebook wants to make sure you've read the article you're about to share. *The Verge* https://www.theverge.com/2021/5/10/22429174/facebook-article-popup-read-misinformation (2021).
85. Brashier, N. M. Fighting misinformation among the most vulnerable users. *Curr. Opin. Psychol.* **57**, 101813 (2024).
86. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
87. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D. G. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychol. Sci.* **31**, 770–780 (2020).
88. Simpson, S. *Fake News: A Global Epidemic Vast Majority (86%) of Online Global Citizens Have Been Exposed to It*. https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic (2019).
89. Seitz, A. & Fingerhut, H. Americans agree misinformation is a problem, poll shows. *AP News* https://apnews.com/article/coronavirus-pandemic-technology-business-health-misinformation-fbe9d09024d7b92e1600e411d5f931dd (2021).
90. Højsgaard, S., Halekoh, U. & Yan, J. The R Package geepack for generalized estimating equations. *J. Stat. Softw.* **15**, 1–11 (2006).
91. Pan, W. Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120–125 (2001).

Acknowledgements

This work was supported by the Knight Foundation, the Office of Naval Research's MURI: Persuasion, Identity, & Morality in Social-Cyber Environments grant N00014-21-12749, Carnegie Mellon University's Graduate small Project Help (GuSH), the Center for Computational Analysis of Social and Organizational Systems (CASOS), and the Center for Informed Democracy and Social-cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors alone. The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank Peter Carragher and Joshua Uyheng for their helpful comments on a draft of this manuscript.

Author contributions

C.K. and K.M.C. conceptualized the study and acquired funding. C.K. wrote the original draft, S.C.P. and K.M.C. contributed to reviewing and editing. C.K. and S.C.P. contributed to data curation, methodology, visualization, and formal analysis. K.M.C. was responsible for project administration and resources.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-93100-7.

Correspondence and requests for materials should be addressed to C.K.