IN PRESS AT THE BRITISH JOURNAL OF PSYCHOLOGY

A Comparison of Prebunking and Debunking Interventions for Implied versus Explicit Misinformation

Li Qian Tay, Mark J. Hurlstone, Tim Kurz, & Ullrich K. H. Ecker

School of Psychological Science, University of Western Australia

Abstract

Psychological research has offered valuable insights into how to combat misinformation. The studies conducted to date, however, have three limitations. First, pre-emptive ("prebunking") and retroactive ("debunking") interventions have mostly been examined in parallel, and thus it is unclear which of these two predominant approaches is more effective. Second, there has been a focus on misinformation that is explicitly false, but implied misinformation that uses literally true information to mislead is common in the real world. Finally, studies have relied mainly on questionnaire measures of reasoning, neglecting behavioural impacts of misinformation and interventions. To offer incremental progress towards addressing these three issues, we conducted an experiment (*N* = 735) involving misinformation on fair trade. We contrasted the effectiveness of prebunking versus debunking and the impacts of implied versus explicit misinformation, and incorporated novel measures assessing consumer behaviours (i.e., willingness-to-pay; information seeking; online misinformation promotion) in addition to standard questionnaire measures. In general, both prebunking and debunking reduced misinformation reliance. We also found that individuals tended to rely more on explicit than implied misinformation both with and without interventions.

Keywords: Misinformation; fake news; refutation; inoculation

A Comparison of Prebunking and Debunking Interventions for Implied versus Explicit Misinformation

Misinformation—information that is ostensibly presented as true but is in fact false or misleading—often influences memory and reasoning even after being corrected. This phenomenon is known as the continued influence effect (Johnson & Seifert, 1994; for a recent meta-analysis, see Walter & Tukachinsky, 2019). The continued influence of misinformation can result in substantial costs across a range of domains, including politics, public health, and climate change (for reviews, see Lewandowsky et al., 2012; Lewandowsky et al., 2017). Although the continued influence literature has been able to provide valuable insights into how to mitigate such costs, it has three weaknesses revolving around the proposed interventions, the types of misinformation investigated, and the outcomes typically assessed. These weaknesses may limit the broad applicability of current recommendations for combatting misinformation. The present study was designed to shed some light on all three issues and offer incremental progress towards addressing them more thoroughly. In what follows, we consider each of these weaknesses in turn before outlining how our study was designed to address them.

Interventions

Interventions to counteract misinformation based on the continued influence literature can be grouped into two broad categories that have, to date, been mostly examined in parallel. The first approach is prebunking, which aims to pre-emptively reduce the persuasiveness of misinformation before it is encoded. One way that prebunking can be achieved is via interventions derived from inoculation theory (Compton, 2013; McGuire, 1964). Inoculation involves (a) a warning of an impending threat to a recipient’s existing beliefs, designed to motivate subsequent counterarguing, and (b) a refutation of an example piece of misinformation that highlights one or more fallacies and techniques typical of misdirection attempts, designed to provide the recipient with counterarguing tools. Inoculations were originally developed to foster resistance against persuasive messages more generally, but have in recent years garnered attention as an effective way to counter misinformation more specifically (e.g., van der Linden et al., 2020). For example, explaining how the tobacco industry recruited spokespeople with only a veneer of expertise to cast doubt on the science linking smoking to lung cancer was shown to shield individuals from the influence of climate-change misinformation involving the same "fake experts" strategy (Cook et al., 2017).

The second approach is debunking, which aims to retroactively reduce reliance on misinformation by correcting it once it has been encoded. Although corrections are rarely completely effective at countering the continued influence of misinformation, several debunking strategies have been identified that make corrections more effective. For instance, corrections can be made more effective by using a trustworthy source (Ecker & Antonio, 2020; Guillory & Geraci, 2013) and by making salient the inconsistency between misinformation and facts, which aids knowledge revision (Kendeou et al., 2019). Other recommendations have included providing additional factual information to explain why the misinformation is false and to provide an alternative interpretation (Johnson & Seifert, 1994; Swire et al., 2017), as well as drawing attention to any misleading strategies that misinformants[1] employ (Cook et al., 2018; MacFarlane et al., 2018). Optimized-debunking formats that take these recommendations into account have been found to be superior in reducing reliance on misinformation compared to

[1] The term *disinformants*—agents that purposefully craft and spread false information to deceive—may be more appropriate here. However, we opted for the more encompassing *misinformants* to accommodate the fact that there could be inadvertent usage of misleading strategies.

standard approaches that often involve only brief or tentative corrections (Ecker et al., 2020; MacFarlane et al., 2020; Paynter et al., 2019).

Although a recent meta-analysis suggested that prebunking may be less effective than debunking, this was based on effect sizes from a small and unbalanced sample of studies that did not directly contrast the two approaches (prebunking: $K = 6$; debunking: $K = 56$; Walter & Murphy, 2018). There are only a few exceptions in which the two approaches are directly compared, which have returned mixed results. On the one hand, Jolley and Douglas (2017) reported significant effects of prebunking but not debunking of vaccine misinformation, and Bolsen and Druckman (2015) found that prebunking was more effective than debunking in countering the politicization of scientific facts. On the other hand, Vraga et al. (2020) found that debunking was successful regardless of whether it focussed on providing factual information or on highlighting the logical flaws of misinformation messages, whereas prebunking was only effective when it focussed on logical flaws. Similarly, Brashier et al. (2021) found debunking was superior to prebunking in terms of improving participants' subsequent ability to discern true and false headlines. Note, however, that both prebunking and debunking in the Brashier et al. study only involved "false" tags that simply indicated that a given headline was false. That is, their interventions did not expose misleading strategies or present accurate information in conjunction with the interventions, unlike studies testing best-practice recommendations. Accordingly, it is unclear which approach, if any, is more effective (and if so, under what circumstances), particularly when more sophisticated misinformation and intervention techniques are adopted.

Misinformation Types

A second limitation of the existing literature is that studies often examine only explicit misinformation—that is, misinformation that can be unequivocally declared false (e.g., a particular factor such as negligence being ruled out as the cause of a fire; Johnson & Seifert, 1994). Yet, implied misinformation—information that is misleading but falls short of literal falsity—is common in the real world, as marketeers, politicians, and individuals with ulterior motives may strive for plausible deniability by using information to mislead that is literally true (e.g., Brown, 2013; Chestnut & Markman, 2018). In fact, even well-intentioned educational campaigns can generate implied misinformation. For example, the American Diabetes Association website states that the answer to the question "*Does eating too much sugar cause diabetes?*" is "*not so simple*", and provides detailed explanations regarding how diabetes may be caused by other factors such as genetics. However, while these other explanations are all true in isolation, they are potentially misleading since they cause individuals to underestimate the causal role of a high-sugar diet (Powell et al., 2020).

Concerningly, implied misinformation may potentially be harder to counteract than explicit misinformation. When exposed to implied misinformation, recipients often generate their own inferences, which can lead to activation of related schemas and richer, more enduring integration of misleading content in memory (Rich & Zaragoza, 2016). Moreover, during attempted correction, explicit misinformation can be directly refuted, making salient the conflict between falsehood and factual correction by fostering their co-activation in memory, which is known to aid knowledge revision (Kendeou et al., 2019). This process may be more difficult with implied misinformation, because individuals may be less able to notice specific points of discrepancy between false and factual information (Rich & Zaragoza, 2016; see also Ecker et al., 2014). Although prebunking may be more effective in this regard, because it aims to pre-emptively reduce the persuasiveness of misinformation and does not depend on knowledge revision at time of correction, no study has empirically tested this possibility.

Outcome Measures

Finally, research exploring the continued influence effect has tended to focus only on the underlying cognitive processes, leaving behavioural impacts underexplored. Indeed, in a typical study, participants are presented with misinformation on a certain topic (e.g., the cause of a fire), which is (or is not) corrected. Then, participants' memory and inferences regarding the topic are assessed using a questionnaire with either open-ended questions (e.g., *What might be a good headline for a report about the fire?*) or rating scales that can be more or less direct (e.g., *Negligence contributed to the fire*; *strongly agree – strongly disagree*; Connor-Desai & Reimers, 2019). This approach has significantly advanced understanding of the continued influence effect, as results show that participants continue to refer to outdated information despite remembering the correction. This means that the reliance is not simply a consequence of failing to notice or remember corrective details (Ecker et al., 2015).

However, one consequence of the above is that the real-world impacts of the continued influence effect and the proposed interventions may, to some extent, rest on the assumption that patterns assessed at the cognitive level will translate well to the behavioural level. Such an assumption should not be taken for granted, as attitudes and beliefs may only weakly and indirectly predict behaviours (e.g., McEachan et al., 2011). Considering that changes to behaviours are often the actual outcomes of interest (e.g., reducing demand for ineffective health products, increasing demand for sustainable products, or reducing the sharing of misinformation on social media platforms), it is critical that this gap be addressed (Hamby et al., 2020; MacFarlane et al., 2020).

The Present Study

The aim of this study was, thus, to investigate the following three questions: Is prebunking or debunking more effective in reducing the impacts of misinformation? Is implied or explicit misinformation more challenging to correct? Will the effects of misinformation and best-practice interventions on measures of cognition extend to measures of behaviour?

To this end, we presented participants with articles containing either implied or explicit misinformation designed to alter consumer behaviours regarding fair-trade products. A no-misinformation group was also included, with participants receiving an article that contained only neutral information about fair trade. We then provided participants with either no intervention, a prebunking treatment prior to reading the misinformation, or a post-exposure debunking of the misinformation. In addition to the standard inference questionnaire, we also examined participants' willingness to purchase products targeted by the misinformation (i.e., fair-trade products), using two measures: bids in hypothetical product auctions, and an end-of-survey question asking if participants would like to receive additional information on where to purchase fair-trade products. Finally, we assessed participants' behaviour when tasked to compose a social-media post (i.e., the extent to which their posts expressed sentiments consistent with the misinformation and would thus promote misinformation when shared online). We tested four pre-registered hypotheses, predicting that (1) both implied and explicit misinformation exposure would increase participants' reliance on misinformation; (2) both prebunking and debunking would be effective in reducing misinformation reliance; (3) implied misinformation would be more difficult to counteract than explicit misinformation; and (4) prebunking would be more effective than debunking for implied misinformation.

Method

The experiment adopted a 2 (misinformation type: implied vs. explicit) × 3 (intervention type: no intervention vs. prebunking vs. debunking) plus control between-subjects design. Ethics approval was granted by [blinded for peer review]. The experiment was pre-registered at https://osf.io/dktnu/?view_only=de4f642566ad4f36aa810ec444e00bfc.

Participants

A total of 735 US-based participants were recruited online through Amazon's Mechanical Turk (MTurk). The minimum eligibility criteria were set at 97% approval rate with at least 5,000 prior tasks completed. As regards to sampling strategy, we first estimated a target of 100 participants per cell based on prior research and resource constraints. We then conducted a simulation-based power analysis using the R package *Superpower* (Lakens & Caldwell, 2021). Results suggested our target sample would produce power of at least 80% at the standard alpha level of $\alpha = .05$, assuming true effect sizes as large as our most conservative estimate (i.e., ordinal interaction with $\eta_p^2 = .02$). The sample size of 735 participants was calculated by extrapolating this to incorporate the control condition and to account for potential missing data. Retaining approximately equal group sizes, participants were randomly allocated to one of the seven conditions.

Materials

Target articles. The articles for control, implied misinformation, and explicit misinformation conditions were all on fair trade. Fair trade as a topic was chosen for three reasons. The first reason relates to the balance between experimental control and real-world relevance. One main benefit of using fictitious event reports in seminal continued influence studies (e.g., the aforementioned warehouse fire) is that such reports allow researchers to minimize the influence of motivational factors and examine continued influence as a cognitive phenomenon, which is important for theorizing. On the other hand, using real-world materials that are associated with consequential behavioural outcomes is important if the goal is to inform application. Fair trade, in our view, provides a middle ground between fictitious event reports and potentially polarizing topics such as climate change and vaccination (e.g., Cook et al., 2017; Jolley & Douglas, 2017). Second, fair trade products are common in the market; thus, individuals will have some conceptual familiarity with fair trade even if they may not hold in-depth knowledge about the movement's efficacy. This allowed us to create plausible misinformation whilst employing willingness-to-purchase measures with products that participants are likely to be familiar with (vs. having participants bid for uncommon or implausible products). Finally, fair trade misinformation was deemed to be relatively benign; for ethical reasons, we opted to not expose participants to potentially harmful misinformation, as this was not required to answer the research question. The control article contained only basic descriptive information about fair trade. The implied-misinformation article built on the no-misinformation article but included the use of misleading strategies. In particular, fake experts and anecdotes were employed to nudge participants towards thinking that fair trade benefits only corrupt middlemen. It did not include literal falsity but instead stopped at the claim being insinuated, leaving participants to draw their own inferences. By contrast, in the explicit-misinformation article, the false claims were stated unequivocally. See Table 1 for examples of critical variations and the Supplement for all presented articles.

Intervention articles. The content of intervention articles was identical in prebunking and debunking conditions; the only difference was the time-point at which the intervention was presented. The intervention was based on best-practice recommendations from the continued influence literature (i.e., it highlighted a trustworthy source, warned of circulating misinformation, highlighted the inconsistency between misinformation and facts, provided additional information, and exposed misleading techniques via examples). It targeted common elements found in both implied and explicit misinformation articles. See Table 2 for details.

Dependent measures.

References-To-Misinformation Questionnaire. We used a point-allocation questionnaire to assess the extent to which participants refer to misinformation in response to inference questions (Connor Desai & Reimers, 2019). The questionnaire included questions such as "*Consumers who buy Fair Trade products are ...*", with the response options "*Supporting the farmers*"; "*Listening to the government*"; "*Being taken advantage of*"; and "*Helping society*". There was a total of five questions, and participants were told that, for each question, they had 10 points to allocate to the options that they prefer. For instance, one could allocate five points to "*Supporting the farmers*", two points to "*Helping society*", and three points to "*Being taken advantage of*", or instead allocate all 10 points to just one preferred option. The points allocated to the misinformation-consistent options (i.e., "*Being taken advantage of*" in the current example) across all questions were summed to create a single references-to-misinformation measure. The measure thus had a minimum score of 0 and a maximum score of 50, with greater scores indicating greater reliance on misinformation.

Willingness-to-Purchase. To assess participants' willingness to purchase fair-trade products, we included two measures. The first measure was a willingness-to-pay measure using an experimental auction (see Kagel & Levin, 2011; MacFarlane et al., 2018). The auction included a sequence of five fair-trade products—blueberries, orange juice, peanut butter, watermelon, and coffee—presented in a randomized order. Each product was displayed on a new screen accompanied by a picture of the product, and participants were asked to imagine that they had been provided with a $4.99 endowment to bid on each product. Participants were asked to bid in dollars and cents, with $b \in (0, 4.99)$. They were told their bids would be compared against a random number $r \in (0, 4.99)$ in each round, and they would purchase the products for rounds in which their bids were equal to or greater than the random number, $b \geq r$, by paying their bid amount, and they would keep the remainder of the hypothetical endowment they chose not to bid. They knew for rounds in which their bid amount was less than the random amount, $b < r$, they would not purchase the product and get to keep the full amount of their hypothetical endowment. Participants were asked to assume for each product they could find a use for it, have the capacity to use it, and there was nothing else preventing them from purchasing the product (e.g., a dislike or allergy). Participants' bids across products were summed to create a single composite measure of willingness-to-pay. The measure had a minimum score of 0 and maximum score of 24.95, with lower scores suggesting greater reliance on misinformation. Willingness-to-pay measures elicited using similar auction formats have previously been shown to be comparable to those elicited under incentivized conditions (i.e., when participants bid with real monetary endowment for physical products; MacFarlane et al., 2020).

The second measure was an end-of-survey question designed to assess willingness-to-seek additional information ("*Thank you, you have completed the questionnaire. Would you like to receive additional information on where you can purchase Fair Trade products online and in stores?*"). Participants were given the option to choose between "*Yes (receive additional* *information)*" or "*No (end survey now)*". This formed the binary willingness-to-seek measure. Participants that chose to receive additional information were provided with a list of fair-trade resources before proceeding to the debrief.

Social-Media Behaviour. The task involved participants composing a *tweet* (i.e., a social media post limited to 240 characters) about fair trade. Participants were reminded to write their posts as if for their actual social-media account, and that their friends, family, and followers would be reading their post. The raw text data was used as input for an exploratory sentiment analysis using the Valence Aware Dictionary and Sentiment Reasoner tool (VADER) to return a sentiment score via the R package *vader*. VADER was chosen for its ability to account for rules that suggest changes in valence strength (e.g., intensifiers and contrastive conjunctions), and its "gold standard" lexicon that has demonstrated comparable or greater accuracy than human raters when applied to short-format social-media data like tweets (Hutto & Gilbert, 2015). The resulting tweet-sentiment scores ranged from -1 to 1. As the misinformation in the current study aimed to denigrate fair trade, lower sentiment scores in one condition relative to others represented greater reliance on misinformation. Participants' tweets were also manually coded 1 (if consistent with the misinformation), 0 (if ambiguous or neutral), and -1 (if inconsistent with the misinformation) to create a propensity-to-promote-misinformation measure. An additional coder coded one-fourth of the sample to assess inter-rater reliability, which was found to be satisfactory (Cohen's $\kappa = .84$). Disagreements between coders were resolved via discussion. See Supplement for examples of tweets and exploratory topic modelling.

Procedure

The experiment was run using Qualtrics software (Qualtrics, Provo, UT). First, participants provided informed consent after reading an ethics-approved information sheet. They were then presented with the article that corresponded to their assigned condition. This was followed by the experimental auction, the inference questionnaire, the social-media task, and the information-seeking question. The study concluded with a debrief that made clear which (if any) statements in the presented article were false, and included web-links to resources for participants should they wish to learn more about fair trade. The study took approximately 13 minutes.

Results

First, to provide an initial test of our hypotheses, as well as a "sanity-check", planned comparisons were conducted on references-to-misinformation, willingness-to-pay, and tweet-sentiment scores. Results are summarised in Table 3. The control condition was contrasted against the implied and explicit misinformation conditions, respectively, to test for the effects of misinformation (top two rows of Table 3); control was contrasted against the prebunking and debunking conditions to test for continued influence effects (bottom four rows of Table 3). Results suggested that both misinformation types increased participants' reliance on misinformation in the expected direction across dependent variables, with explicit misinformation returning larger effect sizes than implied misinformation. Results also suggested that debunking was better able to counteract misinformation than prebunking, as conditions in which participants received the debunking generally returned smaller continued influence effects than when participants received prebunking.

Next, we excluded the no-misinformation control condition to test the remaining hypotheses. For the references-to-misinformation measure, a misinformation type (implied vs. explicit) × intervention type (no intervention vs. prebunking vs. debunking) between-subjects ANOVA was conducted (see Figure 1). This revealed a significant main effect of misinformation type, $F(1, 625) = 8.48$, $p = .004$, $\eta_p^2 = .01$, as well as a significant main effect of intervention type, $F(2, 625) = 53.50$, $p < .001$, $\eta_p^2 = .15$. The interaction was non-significant, $F(2, 625) = 0.78$, $p = .461$, $\eta_p^2 = .002$[2]. Post-hoc tests with Holm-Bonferroni correction revealed that, across misinformation types and relative to no intervention, both debunking and prebunking significantly reduced participants' number of references to misinformation, with $t(625) = -9.60$, $p = < .001$, $d = .89$, and $t(625) = -8.08$, $p = < .001$, $d = .73$, respectively. The difference between debunking and prebunking conditions was not significant, $t(625) = -1.61$, $p = .108$, $d = .28$.

Turning to participants' willingness to purchase fair-trade products, we first analysed the willingness-to-pay data in a two-way ANOVA (see Figure 2). There was no significant main effect of misinformation type, $F(1, 625) = 1.46$, $p = .228$; $\eta_p^2 = .002$, no significant main effect of intervention type, $F(2, 625) = 2.50$, $p = .083$; $\eta_p^2 = .008$, and no significant interaction, $F(2, 625) = 2.40$, $p = .091$; $\eta_p^2 = .008$. For willingness-to-seek information on fair-trade products, we analysed the data with logistic regression models. A likelihood-ratio test revealed that, compared to a null model with only intercept, a model with both misinformation and intervention types as predictors did not significantly improve model fit, $p = .55$, MacFadden's $R^2 = .009$.

[2] For completeness, we note that there was no significant difference between prebunking and debunking when isolating implied misinformation conditions for references-to-misinformation, $t(625) = 1.830$, $p = 0.339$, willingness-to-pay, $t(625) = 0.49$, $p = 1.000$, tweet-sentiment, $t(625) = -1.91$, $p = 0.452$, or propensity-to-promote-misinformation, z-ratio = 2.25, $p = 0.123$. The full set of contrasts can be found in the Supplement.

Finally, for social-media behaviours, a two-way ANOVA on tweet-sentiment scores (see Figure 3) revealed a significant main effect of intervention type, $F(2, 625) = 10.81, p < .001$; $\eta_p^2 = .03$. However, there was no significant main effect of misinformation type, $F(1, 625) = 3.30, p = .070$; $\eta_p^2 = .005$, and no significant interaction, $F(2, 625) = 2.01, p = .135$; $\eta_p^2 = .006$. Collapsed across misinformation types, post-hoc tests with Holm-Bonferroni correction revealed that debunking resulted in more positive sentiments compared to the no-intervention condition, $t(625) = 4.65, p = <.001, d = .46$, as well as compared to the prebunking condition, $t(625) = 2.44, p = .030, d = .24$. There was also a significant difference between prebunking and no-intervention conditions, with prebunking resulting in more positive sentiments, $t(625) = 2.23, p = .030, d = .22$.

We analysed the propensity-to-promote-misinformation measure (see Figure 4) using cumulative-link ordinal regression models that (1) included only an intercept term, (2) additionally included misinformation type as a predictor, (3) additionally included intervention type as a predictor, and (4) additionally included an interaction term. The term for misinformation type significantly improved model fit, $\chi^2(1) = 5.73, p = .016$ (Model 2 vs. Model 1), as did the term for intervention type, $\chi^2(2) = 74.41, p < .001$ (Model 3 vs. Model 2). However, the addition of an interaction term did not result in a statistically significant improvement in fit, $\chi^2(2) = 3.99, p = .136$ (Model 4 vs. Model 3). Post-hoc tests based on the full model with Holm-Bonferroni correction revealed that, when collapsed across intervention types, explicit misinformation resulted in greater propensity-to-promote-misinformation than implied misinformation, z-ratio = 2.44, $p = .015$. When collapsed across misinformation types, both prebunking and debunking significantly reduced propensity-to-promote misinformation relative to no-intervention, z-ratio = -6.29, $p < .001$, and z-ratio = -8.15, $p < .001$, respectively.

Debunking was also found to result in lower propensity-to-promote misinformation relative to prebunking, *z*-ratio = -2.12, *p* = .034.

Discussion

The continued influence literature has, to date, three major limitations. First, as best-practice recommendations on prebunking and debunking have mostly been examined separately, it remains unclear which approach, if any, is more effective. Second, studies have tended to focus only on explicit misinformation, leaving underexplored the effects of implied misinformation that uses literally-true information to mislead. Finally, the outcomes assessed have typically been restricted to questionnaire measures of reasoning and beliefs, even though changes to behaviours are important to real-world impacts. Here, using fair trade as a topic, we examined the relative impact of implied and explicit misinformation, the relative effectiveness of prebunking and debunking in reducing that impact, and incorporated novel behavioural measures in addition to the standard questionnaire measures.

Both implied and explicit misinformation increased participants' reliance on misinformation when responding to the questionnaire and when composing a tweet; both prebunking and debunking were able to reduce misinformation reliance. However, only explicit misinformation reliably impacted willingness-to-pay, and none of the included factors impacted willingness-to-seek additional information. Moreover, implied misinformation was not more difficult to counteract than explicit misinformation, and prebunking was no more effective than debunking for implied misinformation. In fact, results suggest that individuals, on average, rely more on explicit misinformation than implied misinformation, even if the difficulty of correcting both misinformation types did not reliably differ.

Our results add to a growing literature showing the general utility of both prebunking and debunking (e.g., Cook, 2017; Nyhan & Reifler, 2015; Paynter et al., 2019; van der Linden et al., 2020), although, to the best of our knowledge, this study is the first to examine the effects of prebunking using behavioural measures (for effects of debunking on behavioural measures, see Hamby et al. 2020, and MacFarlane et al., 2020). This is also one of the first studies to compare directly pre-emptive and retroactive correction interventions. Results are inconsistent with studies that found prebunking more effective than debunking (e.g., Bolsen & Druckman, 2015; Jolley & Douglas, 2017), but also do not corroborate studies claiming the contrary (e.g., Brashier et al., 2021; Walter & Murphy, 2018). Although debunking appeared more effective than prebunking based on descriptive differences in effect sizes in the present study, the most prudent interpretation is simply that both interventions are indispensable tools in the fight against misinformation. Critically, the present results advance our understanding by showing that the effectiveness of both interventions applies (i) when countering elaborate misdirection (i.e., going beyond headlines to include content with fake experts and anecdotes) with more sophisticated intervention strategies (i.e., following best-practice recommendations of debunking and inoculation), (ii) to both implied and explicit misinformation, and (iii) to not just questionnaire measures but also social-media behaviours.

However, despite showing the potential for literally-true information to mislead, results were contrary to expectations in that we did not find implied misinformation more difficult to correct. This contrasts with the findings of Rich and Zaragoza (2016). However, our interventions were based on best-practice recommendations and thus were likely more efficacious than the brief, single-lined negations used by Rich and Zaragoza. Thus, our interventions might have overcome factors, such as the otherwise difficult task of noticing specific discrepancies between false and factual information, that could have contributed to more enduring effects of implied misinformation. All else equal, the finding that individuals rely more on explicit misinformation may simply reflect the fact that they received and encoded more extensive misleading content.

More generally, our results again highlight the need for continued influence research to go beyond questionnaire measures of memory and reasoning. Indeed, differences between conditions tended to be smaller or not statistically reliable when assessed using behavioural measures (particularly for willingness-to-purchase measures) compared to the questionnaire measures, even if they were numerically in the expected directions. This is consistent with the attitude-behaviour gap (e.g., McEachan et al., 2011), and suggests that the links between specific beliefs (e.g., fair-trade profits only middlemen) and behavioural outcomes such as product demand and social-media behaviours are unlikely to be straightforward. While we refrain from drawing strong conclusions based on the weak evidence obtained, there may also be other factors that affected our behavioural measures (for example, reputation concerns and interestingness of claims have been proposed to shape misinformation-sharing behaviours, beyond accuracy of beliefs; Altay et al., 2020; 2021).

There are clear practical implications. One common reason put forth *against* the correction of misinformation, particularly by official sources or technology platforms, is that attempts to do so may backfire and ironically increase individuals' reliance on misinformation (e.g., Smith, 2017). This is in part due to the worry that repeated mentions of misinformation during exposure and correction can increase its familiarity, as individuals may assume familiar information as facts (e.g., Schwarz et al., 2016). In direct contrast with such concerns, our results suggest that well-designed interventions are integral to combating misinformation even if they repeat the misinformation (see also Swire-Thompson et al., 2020). Importantly, this appears to be the case regardless of the order in which the misinformation and intervention is presented, whether the misinformation is implied or explicit, and across outcome measures.

Nonetheless, the present study has several limitations and possible extensions. First, although our use of multiple novel and exploratory measures is important in offering new insights, what it also means is that specific *p*-values should be interpreted with caution due to the number of tests conducted. While our interpretations were guided by the general patterns and effect sizes, which appear rather consistent across measures, future research is needed to validate and extend current results. This is particularly so for the behavioural tasks, as participants did not bid using real money and did not have their tweets actually communicated to others. It is also unclear if the recruitment of a convenience sample via an online platform with fixed payment rate might have impacted results for the willingness-to-seek measure, as the measure requires interested participants to extend their time commitment. Indeed, although there is some evidence to suggest that results obtained via auctions may be comparable with hypothetical and real monetary incentives (e.g., MacFarlane et al., 2020), there can at times be substantial heterogeneity between the two (e.g., Kanya et al., 2019; Voelckner, 2006), which may apply to all our behavioural tasks. We therefore recommend that future research delve deeper into each individual measure with larger samples (ideally: not relying on convenience sampling or time-constrained participants); investigate the role of incentives (e.g., by offering actual monetary endowments and products); and incorporate more realistic group-based social-media simulations across a range of topics.

Second, the current study employed a single topic of fair trade. Insofar as the processing of fair-trade related misinformation is comparable to the processing of other misinformation (i.e., there is nothing fundamentally different about fair trade as a topic), we do not have reason to believe that the patterns observed will be restricted to the current context. Indeed, current findings can also be interpreted as an extension and generalization of prior research (e.g., Cook et al., 2017, focussed on climate-change misinformation; Jolley & Douglas, 2017, focussed on vaccine misinformation). Nonetheless, future research may consider replications with alternative topics, or a meta-analytical approach to estimate the extent to which specific content may account for variations in misinformation reliance, particularly across implied and explicit misinformation types and different outcome measures.

Third, we did not explore how individual-differences factors may hinder (or facilitate) misinformation interventions. Previous research has demonstrated that individuals may share misinformation because of worldviews and in-group signalling (see Mercier, 2020, and Lewandowsky, 2021), and that older adults tend to demonstrate greater continued influence effects, presumably due to cognitive decline (see Swire et al., 2017). Future research can therefore also consider whether demographic factors, particularly political ideology and age, may moderate results.

Finally, we did not account for one main benefit of prebunking, which is that it can protect against multiple misinformation encounters using the same misleading strategies, potentially up to several months (Maertens et al., 2020). Even with identical intervention content, as in the present study, such protective effects may not be possible with debunking. The misinformation being encountered first may influence subsequent processing, in that the intervention may be encoded with the expectation of being a retroactive application (i.e., the focus is on memory updating and knowledge revision). By contrast, without prior misinformation exposure, as in prebunking, the intervention may be encoded with the expectation of being for future application (see also Klein et al., 2010, for a discussion on how future orientation can enhance memory encoding and thus recall). This is an additional reason not to conclude from an eyeballing of present results that debunking is the superior intervention. It would also suggest that, even in debunking, we should perhaps still orient the audience towards future applications where possible. Follow-up research should therefore consider contrasting the generalisability and longevity of debunking against prebunking, by introducing novel misinformation across time delays. Such investigations could manipulate the time interval prior to assessment of cognitive and behavioural outcomes. The manipulations would help map the boundaries of protective effects and be particularly important for policy implications. Naturally, the long-term effects demonstrated in Maertens et al. (2020) may also be a result of greater cognitive effort required of participants due to the active nature of the inoculation used (compared to passive exposure to text- or video-based materials; e.g., Cook et al., 2017; Lewandowsky & Yesilada, 2021). Another line of inquiry may therefore seek to investigate whether debunking interventions that require active participation may also have more lasting effects.

To conclude, existing misinformation research has mainly investigated prebunking and debunking in parallel, focussed on explicit but not implied misinformation, and on cognitive rather than behavioural outcomes. The present study aimed to offer incremental progress towards resolving all three limitations. Contrary to expectations, results suggest that implied misinformation may not be harder to correct than explicit misinformation. Results also suggest that prebunking and debunking are both effective in reducing misinformation reliance, at least in the short-term, regardless of misinformation type and for all outcomes assessed.

References

Bolsen, T., & Druckman, J. (2015). Counteracting the politicization of science. *Journal of Communication, 65*(5), 745–769. https://doi.org/10.1111/jcom.12171

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications, 162*, 113746–. https://doi.org/10.1016/j.eswa.2020.113746

Brashier, N., Pennycook, G., Berinsky, A., & Rand, D. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences, 118*(5). https://doi.org/10.1073/pnas.2020043118

Brown, A. (2013). Understanding pharmaceutical research manipulation in the context of accounting manipulation. *The Journal of Law, Medicine & Ethics, 41*(3), 611–619. https://doi.org/10.1111/jlme.12070

Chestnut, E., & Markman, E. (2018). "Girls are as good as boys at math" implies that boys are probably better: A study of expressions of gender equality. *Cognitive Science, 42*(7), 2229–2249. https://doi.org/10.1111/cogs.12637

Compton, J. (2013). Inoculation theory. *The Sage Handbook of Persuasion: Developments in Theory and Practice*, *2*, 220-237.

Connor Desai, S., & Reimers, S. (2019). Comparing the use of open and closed questions for Web-based measures of the continued-influence effect. *Behavior Research Methods, 51*(3), 1426–1440. https://doi.org/10.3758/s13428-018-1066-z

Cook, J., Ellerton, P., & Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters, 13,* 024018. https://doi.org/10.1088/1748-9326/aaa49f

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE, 12*(5), e0175799–e0175799. https://doi.org/10.1371/journal.pone.0175799

Ecker, U. K. H., & Antonio, L. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition, 49*(4), 631–644. https://doi.org/10.3758/s13421-020-01129-y

Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied, 20*, 323-335. https://doi.org/10.1037/xap0000028

Ecker, U. K. H., Lewandowsky, S., Cheung, C. S. C., & Maybery, M. T. (2015). He did it! She did it! No, she did not! Multiple causal explanations and the continued influence of misinformation. *Journal of Memory and Language, 85,* 101-115. https://doi.org/10.1016/j.jml.2015.09.002

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review, 18*(3), 570-578. https://doi.org/10.3758/s13423-011-0065-1

Ecker, U. K. H., O'Reilly, Z., Reid, J., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *The British Journal of Psychology, 111*(1), 36–54. https://doi.org/10.1111/bjop.12383

Guillory, J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition, 2*, 201-209. https://doi.org/10.1016/j.jarmac.2013.10.001

Hamby, A., Ecker, U. K. H., & Brinberg, D. (2020). How stories in memory perpetuate the continued influence of false information. *Journal of Consumer Psychology, 30*, 240-259. https://doi.org/10.1002/jcpy.1135

Kagel, J. H., & Levin, D. (2011). Auctions: A survey of experimental research, 1995-2010. *Handbook of Experimental Economics*, *2*, 563-637

Kanya, L., Sanghera, S., Lewin, A., & Fox-Rushby, J. (2019). The criterion validity of willingness to pay methods: A systematic review and meta-analysis of the evidence. *Social Science & Medicine (1982), 232*, 238–261. https://doi.org/10.1016/j.socscimed.2019.04.015

Kendeou, P., Butterfuss, R., Kim, J., & Van Boekel, M. (2019). Knowledge revision through the lenses of the three-pronged approach. *Memory & Cognition, 47*, 33-46. https://doi.org/10.3758/s13421-018-0848-y

Klein, S., Robertson, T., & Delton, A. (2010). Facing the future: Memory as an evolved system for planning future acts. *Memory & Cognition, 38*(1), 13–22. https://doi.org/10.3758/MC.38.1.13

Lakens, D., & Caldwell, A. (2021). Simulation-based power Analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1), 251524592095150. https://doi.org/10.1177/2515245920951503

Lewandowsky, Ecker, U. K. H., Seifert, C., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*, 106-131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the post-truth era. *Journal of Applied Research in Memory and Cognition, 6*, 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

MacFarlane, D., Hurlstone, M. J., & Ecker, U. K. H. (2018). Reducing demand for ineffective health remedies: Overcoming the illusion of causality. *Psychology & Health, 33*, 1472-1489. https://doi.org/10.1080/08870446.2018.1508685

MacFarlane, D., Hurlstone, M. J., & Ecker, U. K. H. (2020). Countering demand for ineffective health remedies: Do consumers respond to risks, lack of benefits, or both? *Psychology and Health.* https://doi.org/10.1080/08870446.2020.1774056

MacFarlane, D., Tay, L. Q., Hurlstone, M. J., & Ecker, U. K. H. (2020). Refuting spurious COVID-19 treatment claims reduces demand and misinformation sharing. *Journal of Applied Research in Memory and Cognition.* https://doi.org/10.1016/j.jarmac.2020.12.005

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology. Applied, 27*(1), 1–16. https://doi.org/10.1037/xap0000315

McEachan, R., Conner, M., Taylor, N., & Lawton, R. (2011). Prospective prediction of health-related behaviours with the theory of planned behaviour: A meta-analysis. *Health Psychology Review, 5*, 97-144. https://doi.org/10.1080/17437199.2010.521684

Nyhan, B., & Reifler, J. (2015). Estimating fact-checking's effects: Evidence from a long-term experiment during campaign 2014. Unpublished manuscript. Retrieved from http://www.americanpressinstitute.org/wp-content/uploads/2015/04/Estimating-Fact-Checkings-Effect.pdf

Paynter, J., Luskin-Saxby, S., Keen, D., Fordyce, K., Frost, G., Imms, C., … & Ecker, U. K. H. (2019). Evaluation of a template for countering misinformation—Real-world autism treatment myth debunking. *PLOS ONE, 14*, e0210746. https://doi.org/10.1371/journal.pone.0210746

Powell, D., Bian, L., & Markman, E. (2020). When intents to educate can misinform: Inadvertent paltering through violations of communicative norms. *PLOS ONE, 15*(5), e0230360–e0230360. https://doi.org/10.1371/journal.pone.0230360

Rich, P., & Zaragoza, M. (2016). The continued influence of implied and explicitly stated misinformation in news reports. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 42*(1), 62–74. https://doi.org/10.1037/xlm0000155

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy, 2*(1), 85–95. https://doi.org/10.1353/bsp.2016.0009

Smith, J. (2017, Dec 21). Designing against misinformation. *Facebook Design.* https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition, 9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 43*, 1948-1961. https://doi.org/10.1037/xlm0000422

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology, 11.* https://doi.org/10.3389/fpsyg.2020.566790

Voelckner, F. (2006). An empirical comparison of methods for measuring consumers' willingness to pay. *Marketing Letters, 17*, 137–149. https://doi.org/10.1007/s11002-006-5147-x

Vraga, E., Kim, S., Cook, J., & Bode, L. (2020). Testing the effectiveness of correction placement and type on instagram. *The International Journal of Press/politics, 25*(4), 632–652. https://doi.org/10.1177/1940161220919082

Walter, N., & Murphy, S. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs, 85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Table 1

Examples of Critical Variations Across the Target Articles

Target Article	Excerpt
Control	"Traders need to guarantee farmers a minimum price regardless of market value, and consumers need to show their support by purchasing products with Fair Trade labels, usually at a higher price than regular products."
Implied Misinformation	"Recent evidence suggests that farmers often incur additional time and costs to obtain and comply with Fair-Trade certification—but not all the extra money that consumers spend goes to them. Why do you think large organizations are competing to be the middlemen for Fair Trade? Who benefits the most? Something needs to change." "I have an acquaintance who's an executive at a Fair Trade intermediary organization. Since starting the job, she has bought a few houses and even a beachfront property in Cape Cod."
Explicit Misinformation	"Recent evidence suggests that farmers often incur additional time and costs to obtain and comply with Fair-Trade certification—but not all the extra money that consumers spend goes to them. Why do you think large organizations are competing to be the middlemen for Fair Trade? Who benefits the most? It's not the producers, that's for sure—Fair Trade is a big rip-off and up to 95% of the extra money is just soaked up by the bureaucracy and intermediary parties. Something needs to change." "I have an acquaintance who's an executive at a Fair Trade intermediary organization. Since starting the job, she has bought a few houses and even a beachfront property in Cape Cod. It's the middlemen who profit, while the farmers supplying the actual goods barely get anything."

Table 2

Examples of Techniques Used Across the Intervention Articles

Intervention Technique	Excerpt
Highlight a trustworthy source	"Dr Alex Davis, a researcher based at the University of Michigan, …"

Warn about misinformation	"…, misinformation from fake experts and anecdotes has been circulating, claiming that Fair Trade is a scam."
Make salient the discrepancy between misinformation and facts, and provide additional information	"One of the false claims about Fair Trade is that most money gets taken by middlemen. That is simply not true. Of course, no system is perfect, and there may be individual cases where Fair Trade delivers no huge benefit. However, we need to look at the actual evidence! In a recent large-scale study of producers …"
Expose misleading techniques	"Fake experts are commentators who either lack relevant expertise or are biased because they have a vested interest. For example, in the 1930-40s, the tobacco industry engaged physicians for advertising purposes. They continued to mislead the public … Now, this technique is being used again to cast doubt on Fair Trade by industry-funded think tanks with vested interests."
	"Anecdotes are personal experiences or isolated examples that can appear convincing at first glance, but can misrepresent the broader evidence. People love stories and anecdotes because they are relatable, so we are often influenced by them even though we shouldn't be."

Table 3

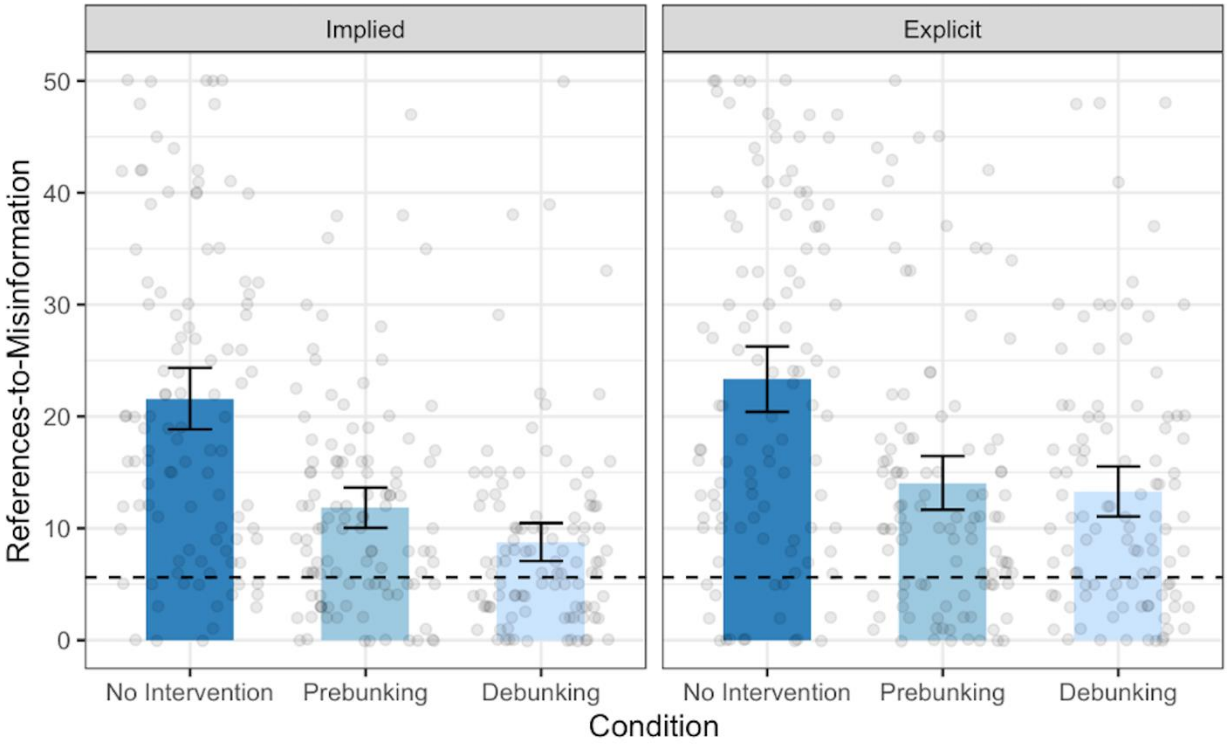
Planned Comparisons for References-To-Misinformation, Willingness-To-Pay, and Tweet-Sentiment

Contrast	Ref.-To-Misinformation			Willingness-to-Pay			Tweet-Sentiment		
	t	*p*	*d* (95% CI)	*t*	*p*	*d* (95% CI)	*t*	*p*	*d* (95% CI)
Effect of Implied Misinformation	10.03	< .001	1.39 (1.11, 1.67)	-1.59	.113	-0.22 (-0.49, 0.05)	-2.69	.007	-0.37 (-0.64, -0.10)
Effect of Explicit Misinformation	11.22	< .001	1.54 (1.26, 1.82)	-2.67	.008	-0.37 (-0.64, -0.10)	-5.47	< .001	-0.75 (-1.02, -0.48)
CIE of Implied Misinformation (Prebunking)	3.93	< .001	0.54 (0.27, 0.81)	-0.60	.548	-0.08 (-0.35, 0.19)	-2.43	.015	-0.33 (-0.60, -0.06)
CIE of Implied Misinformation (Debunking)	1.97	.049	0.27 (0.00, 0.55)	-1.08	.282	-0.15 (-0.42, 0.12)	-0.45	.653	-0.06 (-0.34, 0.21)
CIE of Explicit Misinformation (Prebunking)	5.30	< .001	0.73 (0.46, 1.00)	-2.59	.010	-0.36 (-0.63, -0.09)	-2.48	.013	-0.34 (-0.62, -0.07)
CIE of Explicit Misinformation (Debunking)	4.77	< .001	0.67 (0.39, 0.94)	-0.08	.937	-0.01 (-0.29, 0.26)	-0.87	.385	-0.12 (-0.40, 0.15)

Note. Ref. = References; CIE = Continued Influence Effect. Degrees of freedom for all contrasts = 728. The first two contrasts compare misinformation conditions against the control condition, which served as a counterfactual “baseline”; the last four contrasts compare correction conditions against control.

Figure 1

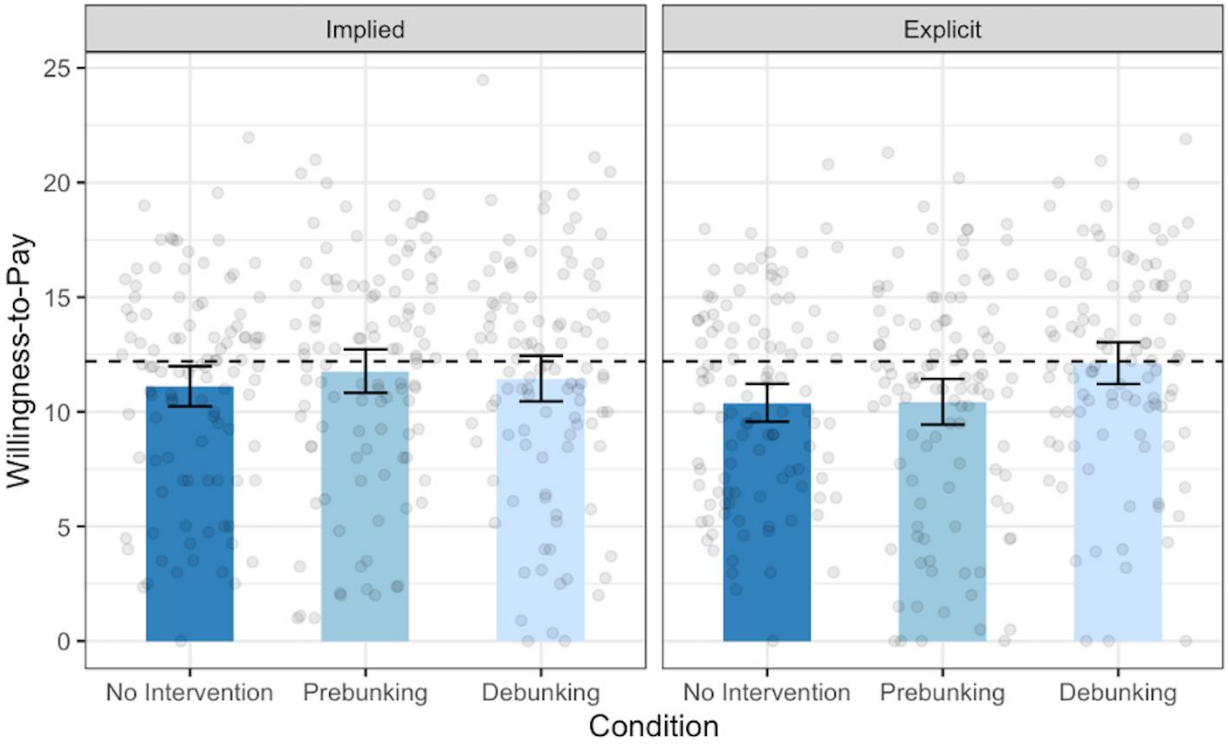
References-to-Misinformation Across Intervention and Misinformation Types



Note. Error bars represent 95% confidence intervals. Dashed line represents mean of the no-misinformation control condition.

Figure 2

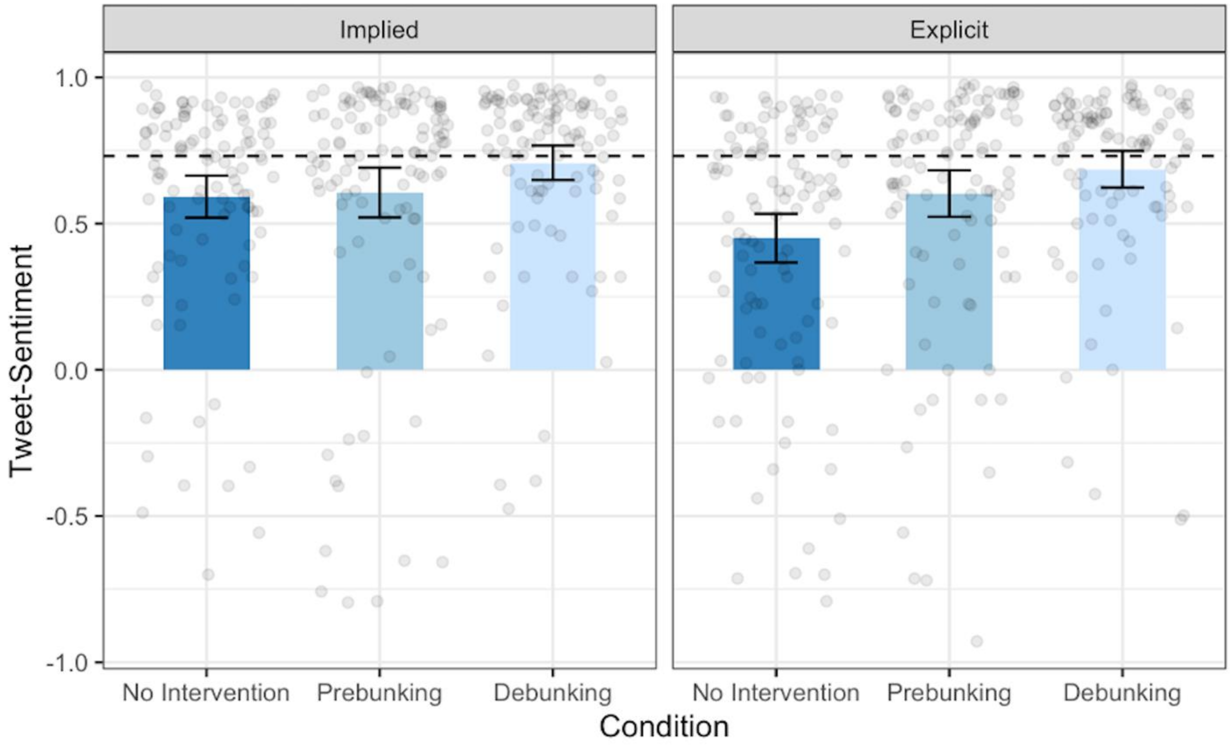
Willingness-to-Pay Across Intervention and Misinformation Types



Note. Error bars represent 95% confidence intervals. Dashed line represents mean of the no-misinformation control condition.

Figure 3

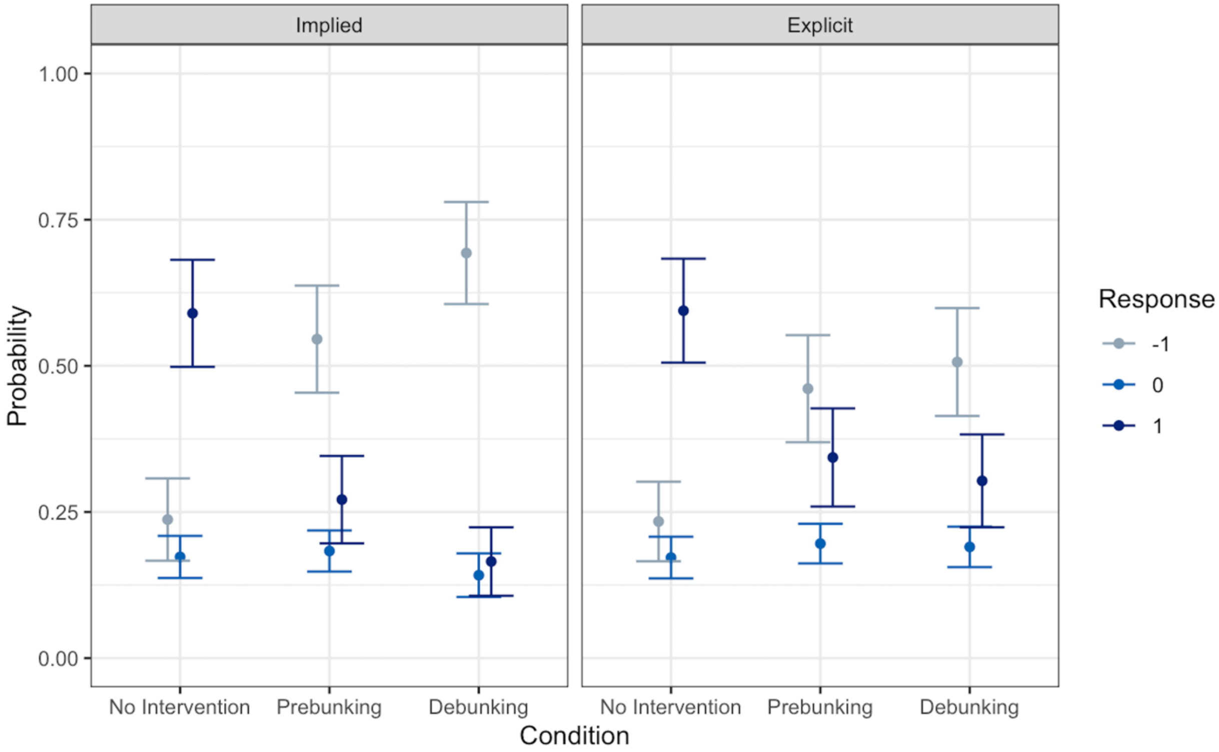
Tweet-Sentiment Across Intervention and Misinformation Types



Note. Error bars represent 95% confidence intervals. Dashed line represents mean of the no-misinformation control condition.

Figure 4

Predicted Probabilities of Response Options Across Intervention and Misinformation Types



Note. Responses were coded 1 if consistent with the misinformation, 0 if ambiguous or neutral, and -1 if inconsistent with the misinformation. Error bars represent 95% confidence intervals.