

The role of (trust in) the source of prebunks and debunks of misinformation. Evidence from online experiments in four EU countries

Authors

Hendrik BRUNS^{*}, François J. DESSART[†], Michał KRAWCZYK[‡], Stephan LEWANDOWSKY[§], Myrto PANTAZI^{**}, Gordon PENNYCOOK^{††}, Philipp SCHMID^{‡‡}, Laura SMILLIE^{§§}

Abstract

Misinformation surrounding crises poses a significant challenge for public institutions. Understanding the relative effectiveness of different types of interventions to counter misinformation and understanding which segments of the population are most or least receptive to them, is crucial. We conduct a preregistered online experiment involving 5,228 participants from Germany, Greece, Ireland, and Poland. Participants were exposed to misinformation on climate change or Covid-19. In addition, they were pre-emptively exposed to a prebunk, warning them of commonly used misleading strategies, before encountering the misinformation, or a debunking intervention afterward. The source of the intervention (i.e. the European Commission) was either revealed or not. Findings show that both interventions effectively change the four outcome variables in the desired direction in almost all cases, with debunks sometimes being more effective than prebunks. Moreover, revealing the source of the interventions does not significantly impact their overall effectiveness. Although one case of undesirable effect heterogeneity – debunks with revealed source were less effective in decreasing credibility of misinformation for people with low trust in the European Union – was observed, the results mostly suggest that the European Commission, and possibly other institutions, can confidently debunk and prebunk misinformation regardless of the trust level of its recipients.

Keywords: misinformation; experiment; debunking; prebunking, inoculation

^{*} Corresponding author. E-Mail: hendrik.bruns@ec.europa.eu, European Commission Joint Research Centre, Rue du Champ de Mars 21, 1050 Brussels, Belgium. Tel. : +32 229-58350.

[†] European Commission Joint Research Centre, Calle Inca Garcilaso 3, 41092 Sevilla, Spain.

[‡] European Commission Joint Research Centre, Rue du Champ de Mars 21, 1050 Brussels, Belgium.

[§] University of Bristol, Priory Road, Bristol, BS8 1TU, United Kingdom.

^{**} Center for Social and Cultural Psychology, Université Libre de Bruxelles, 50 Avenue Franklin Roosevelt, 1050 Brussels, Belgium.

^{††} University of Regina, 3737 Wascana Parkway, Regina, SK S4S 0A2, Canada.

^{‡‡} Universität Erfurt, Nordhäuser Str. 63, 99089 Erfurt, Germany.

^{§§} European Commission Joint Research Centre, Rue du Champ de Mars 21, 1050 Brussels, Belgium.

Introduction

Misinformation is prevalent in various crises, such as the climate crisis and Covid-19. Climate change misinformation includes doubts about human involvement in global warming, denial of its existence, and rejection of the scientific consensus¹. Similarly, Covid-19 has been accompanied by misinformation from the start, including narratives that question its existence, downplay its severity, promote unproven remedies, and cast doubt on the efficacy of vaccination².

In addition to the threat posed by crises themselves, misinformation around crises threatens societies and increases for difficulty of public institutions to address these crises. Believing in Covid-19 misinformation can discourage protective behaviour^{3,4}, including vaccination⁵, with potentially life-threatening consequences⁶. Exposure to climate change misinformation decreases pro-social behaviour and acceptance of scientific facts⁷. Addressing and managing misinformation has therefore become a crucial component of an effective crisis-response, particularly when it jeopardizes public discourse, institutional integrity, and public health⁸.

Public institutions have access to science-based interventions to combat misinformation, including debunks and prebunks⁹. Debunks involve exposing and refuting false information with credible sources *after* exposure to misinformation^{10–12}. Prebunks, on the other hand, proactively warn individuals about misinformation *before* exposure, refute often used erroneous arguments, and explain the strategies commonly used in spreading false information^{13–22}.

Both prebunking and debunking interventions have been found to be effective in reducing the threat of misinformation^{11,13,14,17,21–26}. This paper addresses four main gaps in the literature, with four corresponding research questions. First, although exceptions exist^{27–29}, prebunking and debunking interventions have been typically investigated separately, leading to scarce evidence on their relative effectiveness. In this paper, we compare the relative effectiveness of the two approaches, providing valuable insights to enable public institutions and policymakers to select the most efficient interventions in times of crises.

Second, existing evidence on the effect of the source of these interventions on their effectiveness is still inconclusive. People evidently consider the source when assessing the credibility of information^{30,31} and misinformation^{32,33}. They appear to do so also for debunks^{33–36}. However, the role of source information for prebunks is unclear. This paper aims to uncover whether revealing the source of an intervention against misinformation modifies its effectiveness. We use the European Commission (referred to as ‘EC’) as the source of the intervention in the experiment due to the major role that this institution played in the fight against Covid-19 misinformation in the European Union (EU)³⁷.

Third, people’s trust in the source of misinformation-countering interventions may be fundamental to their success, and yet there is a lack of evidence looking into this. This study examines whether the effectiveness of misinformation-countering interventions depends on recipients’ levels of trust in the EU (i.e. the source of our interventions). Trust in the EU is a commonly assessed measure, used here as an indicator of trust in the European Commission, which is the common source of public campaigns like those aiming at combatting misinformation. Source credibility may matter more to some people than to others and recent findings suggest that tailored interventions taking perceived credibility into account may be worthwhile³⁸.

Fourth, much of the available evidence is based on US samples. For instance, a debunking meta-analysis from 2018 consisted of almost 70% studies with a North American sample²³. This study uses a wide non-US, multi-country sample to reach a wide generalisability of the findings. More specifically, our experiment involved 5,228 participants from four European countries (Germany, Greece, Ireland and Poland). In addition, this study examines the efficacy of prebunking and debunking interventions for misinformation on two topics rather than just one (i.e. Covid-19 and climate change) and includes comprehensive outcome variables, capturing not only self-reported beliefs but also intentions to share misinformation online and offline, with public or close contacts, and differentiating between endorsing and condemning its content as motives for sharing.

Methods and Materials

Participants

This study was run in October 2022 with a total of N=5,228 participants who completed the experiment (Germany: n=1,311; Greece: n=1,313; Ireland: n=1,296; Poland: n=1,308). All participants who finished the survey were included in the analyses, consistent with the preregistration.

The sampling process involved quotas based on age, gender, and geographic region (NUTS regions) to ensure a representative sample of each country's population. Among the respondents, 52.22% identified as female, 46.89% as male, and the remaining respondents chose none of those. The age distribution was as follows: 9.7% were between 18 and 24, 15.61% between 25 and 34, 18.1% between 35 and 44, 18.06% between 45 and 54, 25.84% between 55 and 64, and 12.69% were above 65 years old. For detailed regional spread and sample characteristics by country please refer to Table S-12 – Table S-16 in the Supplementary Material.

Power analysis

A power analysis was conducted using data from a pilot experiment consisting of 875 observations (more information on the pilot experiment is provided below). This calculation assumed a 5% significance threshold and a two-tailed z-test from a logistic regression. The required number of observations is 1,300 participants from each country.

Preregistration and ethical approval

The preregistration is available at [aspredicted.org](https://aspredicted.org/blind.php?x=5XH_2QP) under https://aspredicted.org/blind.php?x=5XH_2QP. The experiment was reviewed and cleared by the Ethics Committee of the Faculty of Economic Sciences, University of Warsaw.

Exclusion criteria

Following the preregistration, only participants who did not complete the full survey were excluded from the dataset. A total of N=5,665 observations were removed, which includes participants who were screened out due to quota requirements. Among the exclusions, at least N=2,305 participants (21.16%) voluntarily dropped out, while N=3,361 individuals did not proceed beyond the screening stage. Although we cannot distinguish between participants screened out by us and those who dropped voluntarily at the screening stage, our experiment monitoring indicates that the majority were screened out by us.

Recruitment and experimental treatments

The experiment was conducted online using Limesurvey. Participants were recruited and paid a fixed amount by online panel provider Ipsos NV. Participants were randomly assigned to one of five treatment groups. Participants read either a prebunking message (prebunk), a debunking message (debunk), or no message (control). Both the prebunk and debunk messages were further subdivided according to the information on the source responsible for their implementation. There was either no information (no source) or information that the European Commission implemented either intervention (EC source). Thus, the design was a 2 (intervention: prebunks vs. debunk) x 2 (intervention source: no source vs. European Commission) + 1 (control) between-subjects design. Furthermore, we introduced between-subjects variation regarding the topic of misinformation and the specific misleading article. Thus, the factorial design was extended to a 2 (intervention: prebunks vs. debunk) x 2 (intervention source: no source vs. European Commission) x 2 (misinformation topic: climate change vs. Covid-19) x 3 (misinformation claim: claim 1 vs. claim 2 vs. claim 3) + 1 (control) between-subjects design. Both the topic and misinformation claim factors serve as robustness checks rather than treatment factors for which we are interested in the treatment effect. Consequently, our main analyses aggregate over these two factors (however, see Supplementary Material for a discussion of differences by content).

Experimental materials

The Supplementary Material (Experimental materials) contains the texts used for interventions and misleading articles, along with examples of their presentation. The prebunks and debunks were designed to be nearly identical, with debunks including all the information from prebunks and additional details specific to the misinformation addressed. Debunks informed those who have encountered specific misinformation *after* the fact, while prebunks were more general and *preceded* encounters with misinformation. By designing both prebunks and debunks in a similar way, we can compare the effectiveness of both interventions.

As regards the misleading articles, there were three possible claims for climate change and Covid-19, respectively. These six claims were selected from a set of 17 claims: eight on Covid-19 and nine on climate change (more on how these were selected below). Apart from specific claims and pictures, the misleading articles were identical. To create the articles, a misinformation claim was combined with a catchy headline, picture, and teaser text. Common misinformation techniques were employed in the generic text, and the article was edited to resemble a typical online news item, including a blurred date and author information. The articles used common misinformation techniques to enhance their credibility, including appeals to emotions, morality, and claims of absolute truth. They also employed strategies to undermine contrary claims, such as questioning the credibility and morality of experts, or alleging the existence of a conspiracy.

The selection of three Covid-19 and three climate change claims involved two pre-tests (different from the pilot described below). The first pre-test, conducted in May 2022 in Germany, Greece, and Poland, had 301 participants rating a random set of four candidate articles out of eight on Covid-19. The second pre-test took place in September 2022 in Germany, Greece, Ireland, and Poland, with 416 participants rating a random selection of four articles out of nine for climate change claims. The original set of 17 claims was sourced from real claims found online (Climate change: Skeptical Science Website^{***}; Covid-

^{***} <https://skepticalscience.com/argument.php>

19: ESOC COVID-19 Misinformation Dataset⁺⁺⁺). Participants rated each misleading article's credibility, indicated their intentions to share it, and in case they wanted to share it, their reason to do so (more detail on the outcome variables is provided below). The respective three articles were selected by ranking the articles from highest to lowest for each outcome variable separately and then counting the number of times the respective article had been in the top three of articles for each outcome variable. The 6 final claims that were selected for the main experiment were (1) "It hasn't warmed since 1998"; (2) "There is no scientific consensus on climate change"; (3) "Climate models are unreliable"; (4) "The Covid-19 vaccine does not work"; (5) "The Covid-19 vaccine has not been properly tested in clinical trials"; (6) "The Covid-19 vaccine is dangerous". Examples for the used articles are shown in the Supplementary Material (Experimental materials). Participants in pre-tests did not participate in the main experiment.

Experimental procedure

After reading an introduction and explanation for the experiment, participants followed a specific sequence based on their assigned intervention treatment. In the prebunk condition, participants received the prebunking message before reading the misleading article. In the debunk condition, participants read the debunking message after reading the misleading article. The control condition involved participants only reading the article (for the specific sequence and elements contained therein see Table 1).

After receiving the intervention and reading the misleading article, participants answered three groups of questions in the following order: First, participants stated their belief in the respective misinformation claim on a 5-Point Likert scale ranging from 'strongly disagree' to 'strongly agree'. For example, participants who had read the misleading article claiming that "it hasn't warmed since 1998" indicated their agreement with this very statement. Second, participants reported their intentions to engage with the misinformation article and, conditional on their response, gave the reasons for their intention. For the former, they indicated their intentions to (a) share the article *online* with *people who were close to them*; (b) share the article *online* and *publicly*; (c) *talk face-to-face* about the article to *people who were close to them*; and (d) *talk face-to-face* about the article *publicly*. Participants indicated their agreement on 5-point Likert scales with options 'not at all', 'a little', 'neither a little nor a lot', 'much' and 'very much'. Respondents who chose anything else than "not at all" at least once were asked about their reason for wanting to engage with the article. Participants provided their response on a 5-Point Likert scale ranging from 'To express that I totally disagree with it' to 'To express that I totally agree with it'. Third, participants indicated their perceptions of credibility of the article. Specifically, they assessed credibility on four dimensions, using 5-point semantic differential scales⁶⁶. These dimensions assessed credibility with respect to accuracy ('inaccurate' – 'accurate'), believability ('unbelievable' – 'believable'), factuality ('opinionated' – 'factual') and trustworthiness ('untrustworthy' – 'trustworthy'). All outcome variables were forced choice, with no 'I don't know' or 'Do not want to say' options.

After responding to these questions, participants entered a post-experimental questionnaire. Most importantly, they reported their levels of trust in the EU. Specifically, they answered the question "How much trust do you have in the European Union?" on a 10-point scale ranging from 'I do not trust it at all' to 'I trust it completely'. After the questionnaire, all participants were debriefed. The other questions related to trust in the national government of the respondent, general trust, agreement with EU-specific statements, perceptions of the prebunk or debunk, the perceived source of the prebunk or debunk, and

⁺⁺⁺ <https://esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset>

further general questions related to perceptions of misinformation. See the Supplementary Material for questionnaire questions and the debriefing message.

Table 1. Experimental sequence for different treatments.

Treatment	Introduction	Prebunk	Misinformation	Debunk	DVs	Questionnaire	Debriefing
Control	YES	NO	YES	NO	YES	YES	YES
Prebunk	YES	YES	YES	NO	YES	YES	YES
Debunk	YES	NO	YES	YES	YES	YES	YES

Notes: Indicates whether a specific component of the experiment (column) occurs in the respective intervention treatment (rows).

Comprehension checks

Participants were presented with two comprehension check questions, one after being exposed to the prebunk or debunk, and another after reading the misleading article. These questions assessed understanding of the intervention and the misinformation. If participants answered a question incorrectly or left it unanswered, they were instructed to review the corresponding text (prebunk, debunk, misinformation) before proceeding.

Pilot experiment

The pilot experiment was conducted in May 2022, with a total of N=875 participants completing it (Germany: n=293; Greece: n=282; Poland: n=300). All observations were included in the analysis. Participants were sampled based on quotas to ensure a sample representative of each country's public, considering age, gender, and geographic region (NUTS regions). The participant breakdown was 51.31% female, 48.69% male, with age distributed as follows: 29.87% were between 18 and 34, 22.75% between 35 and 44, and 46.79% between 45 and 64 years old (0.58% did not provide a response). The pilot aimed to test the initial design, identify potential improvements, and generate initial estimates for effect sizes to inform power analyses for the main experiment. It led to changes in the experiment's sequencing and the inclusion of a control group. The pilot experiment focused on debunk interventions for Covid-19 misinformation.

Analysis

The four main variables were analysed according to the preregistration as follows: Agreement with the claim was analysed using an ordered logit model with the ordered response variable. Credibility assessments were analysed using an ordinary least squares model, summing the four credibility responses as the dependent variable. Behavioural intentions were analysed using two binary logistic models, dichotomizing the ordered variable to represent whether respondents expressed intentions to circulate the misleading article and indicated doing so to express (dis-)agreement or total (dis-)agreement, zero otherwise.

For all main hypothesis tests (i.e., the interaction effects), the independent variables included the intervention source, the metric EU trust variable, and their interaction. Heteroskedasticity robust standard errors were used for all model estimations.

Robustness checks

We conducted robustness checks for key analyses, as preregistered. These checks control for age, gender, level of education, country of residence, political ideology, trust in the national government, general trust, a trust index in the EU, need for cognition, frequency of social media use, perceived frequency of misinformation encounter, perceived importance of sharing true information, confidence in identifying misinformation, as well as responses to the comprehension check questions, and a manipulation check regarding the correct identification of the debunk/prebunk source. Analyses incorporating the misinformation topic are conducted separately in the Supplementary Material.

Results

Do debunks and prebunks work?

Compared to the control condition, where no prebunking or debunking intervention was provided, all four interventions significantly and substantially reduced agreement with the misleading article's claim (Figure 1a, with detailed tables in Table S-5 of the Supplementary Material). These effects are substantial, approximately halving the odds of strongly agreeing with the main (false) claims. There is also a significant association (i.e. main effect) between trust in the EU (mean-centred) and agreement: participants with higher trust in the EU were less likely to agree with the claim. A one-standard deviation increase in EU trust had an effect identical to the prebunks.

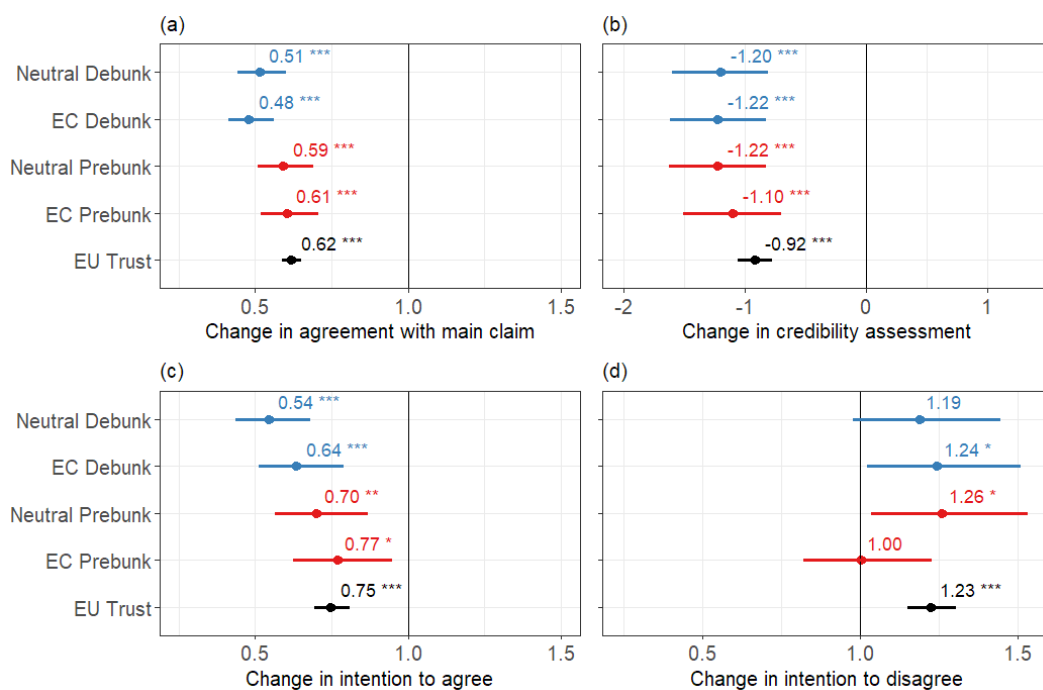


Figure 1. Effects of debunks and prebunks revealing (i.e. EC – European Commission) or not revealing (i.e. neutral) the source of the intervention on the main outcome variables. The y-axis shows the four experimental treatments (with Control as the reference condition) and standardized trust in the EU. The x-axis shows the changes in the four main outcome variables. (a) shows the effects on agreement with the main claim shown in the misleading article from an ordered logistic regression as odds ratios; (b) shows the effects on credibility assessments of the misleading article from a linear OLS regression as linear estimates; (c) shows the effects on intentions to share the misleading article to express agreement with it (i.e. 'intention to agree') from a binary logistic regression as odds ratios; (d) shows the effects on intentions to share the misleading article to express

disagreement with it (i.e. 'intention to disagree') from a binary logistic regression as odds ratios. Effects of debunks are shown in blue, prebunks in red. Bars represent heteroscedasticity-robust 95% confidence intervals. Significance levels: *** <.001, ** <.01, * <.05.

Debunks and prebunks had significant, negative, and meaningful effects on the credibility assessment of the misleading article, regardless of the source (Figure 1b). Again, there is a significant (main) association with trust in the EU. All interventions successfully reduced credibility assessments by more than one point on the credibility scale (which ranges from 0 to 16). Additionally, a one standard deviation increase of trust in the EU was associated with a decrease of almost one point on the credibility rating scale, slightly below the intervention effects.

The interventions significantly decreased participants' intentions to share the misleading article to express their agreement, as shown in Figure 1c. Neutral debunks reduced the odds of intention to share to show agreement by almost half compared to the control (no intervention) treatment, followed by EC debunks with an odds ratio of 0.64. The effects of the two prebunks were also significant. Importantly, the neutral debunk was significantly more effective than the neutral prebunk ($OR=0.78$, $CI_{95}=[0.61-0.99]$, $p=0.038$) and the EC debunk was more effective than the EC prebunk ($OR=0.71$, $CI_{95}=[0.56-0.90]$, $p=0.004$). As for the previous outcome variables, participants with high trust in the EU displayed lower intentions to agree with the misleading article with an effect size similar to the EC prebunk.

For participants' willingness to share or discuss the misleading article to express disagreement (Figure 1d), the effects are less pronounced than for the previous outcomes. Two interventions increased the likelihood of such an intention, but the effects are weaker than for other outcome variables. In particular, the EC debunk and the neutral prebunk slightly increased the likelihood of participants wanting to share the misleading article to express disagreement. However, the neutral debunk and the EC prebunk did not have a significant impact, although the effect of the neutral debunk becomes significant when controlling for all covariates specified in the preregistration ($OR=1.39$, $CI_{95}=[1.08-1.79]$, $p=0.01$). As expected, higher trust in the EU was associated with a higher likelihood of sharing the misleading article to disagree with it, with similar intensity to the effective interventions.

The main effects presented above remain robust when accounting for all specified covariates in the preregistration (see Table S-17, Supplementary Material). Robustness checks involved performing main analyses with additional controls for subject characteristics, responses to comprehension check questions, correct identification of the intervention source, and the misinformation topic (more detail is provided in the Method section).

Do debunks or prebunks work better?

Controlling for source-reveal and trust in the EU, we observe two significant differences between debunks and prebunks (see Figure 2). Firstly, debunks are more effective than prebunks in reducing agreement with the main claim ($OR=0.83$, $CI_{95}=[0.74-0.92]$, $p=0.001$). Secondly, debunks are more effective in decreasing the likelihood of sharing to express agreement with the false claim ($OR=0.701$, $CI_{95}=[0.68-0.95]$, $p=0.008$). However, there are no significant differences regarding the other two outcome variable, i.e., credibility assessment ($E=0.05$, $CI_{95}=[-0.23-0.33]$, $p=0.725$) and intentions to share to express disagreement with the misleading article ($OR=0.92$, $CI_{95}=[0.80-1.05]$, $p=0.218$), where both interventions perform equally well. The latter non-significant effect may be partly due to floor effects. Overall, these results suggest a (very) small advantage of debunks with respect to prebunk to address misinformation.

Does revealing the source change the effectiveness of debunks and prebunks?

Figure 1 already compared the effects of both EC and neutral debunks and prebunks. To further explore this, Figure 2 presents the effects of EC-source compared to neutral source (i.e. no source), controlling for intervention and trust in the EU. Detailed results can be found in Table S-6 in the Supplementary Material. As can be seen, the estimates for EC source are non-significant across all outcome variables, indicating that EC-source does not significantly alter the effectiveness of the interventions in influencing the main outcome variables, on average.

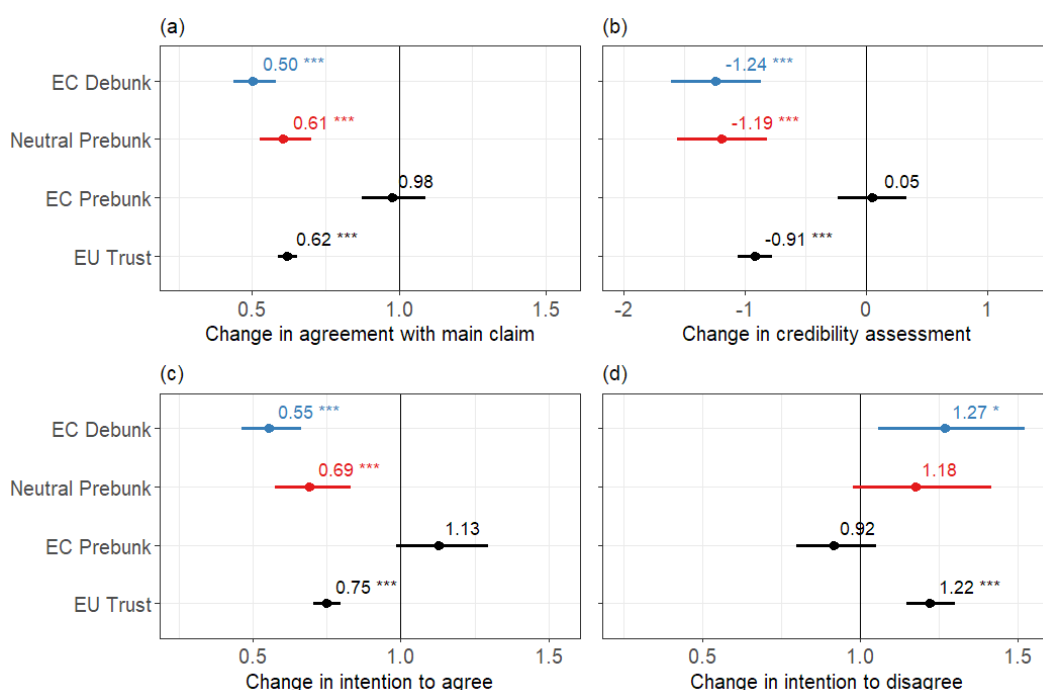


Figure 2. Effects of debunks, prebunks, and revealing the European Commission (i.e. EC) as intervention source on the main outcome variables. The y-axis shows the interventions (with Control as the reference condition), the EC as source of the intervention (vs. neutral, i.e. no source), and standardized trust in the EU. The x-axis shows the changes in the four main outcome variables. (a) shows the effects on agreement with the main claim shown in the misleading article from an ordered logistic regression as odds ratios; (b) shows the effects on credibility assessments of the misleading article from a linear OLS regression as linear estimates; (c) shows the effects on intentions to share the misleading article to express agreement with it (i.e. ‘intention to agree’) from a binary logistic regression as odds ratios; (d) shows the effects on intentions to share the misleading article to express disagreement with it (i.e. ‘intention to disagree’) from a binary logistic regression as odds ratios. Effects of debunks are shown in blue, prebunks in red. Bars represent heteroscedasticity-robust 95% confidence intervals. Significance levels: *** < .001, ** < .01, * < .05.

Does the effect of revealing the source vary based on people’s trust in the source?

We tested the pre-registered interaction effects between the source of the debunk/prebunk (i.e. the European Commission) and participants' reported trust in the EU (Figure 3). The detailed values can be found in Table S-7 and Table S-8 in the Supplementary Material. The regression analyses incorporate an interaction term between the treatment variable and mean-centred EU trust. Therefore, the predictors for the treatment variable represent its effect for people with average levels of trust in the EU, while the EU trust variable indicates the association between trust in the EU and the outcome variable for individuals in the reference treatment group (receiving neutral debunks or prebunks).

The top row of Figure 3 (panels (a) and (b)) illustrates the effects of EC interventions compared to neutral interventions on belief in the claim and credibility assessments of the misleading article for both debunks and prebunks, for different levels of EU trust. Significant interactions are observed in these cases. Specifically, as trust in the EU increases, EC debunks were more effective in reducing agreement with the main claim compared to neutral debunks. This effect is prominent among respondents with high EU trust. However, no such interaction effect is observed for prebunks. The interaction effect for debunks diminishes and becomes insignificant when all preregistered covariates are included ($OR=0.87$, $CI_{95}=[0.71-1.07]$, $p=0.192$). Conditional effects (panel (a-ii)) are not robust to controlling the false discovery rate, which is recommended when conducting multiple hypothesis tests at different levels of the conditioning variable (in our case: level of trust in the EU)³⁹.

In Figure 3b, a more pronounced interaction effect is observed for perceived credibility. As trust in the EU increased, the EC debunk was more effective than the neutral debunk in reducing perceived credibility of the misleading article. This effect is evident in panel (b-ii), where EC source decreases perceived credibility of the misleading article for the debunking intervention among individuals with high trust in the EU but is counterproductive among individuals with low trust in the EU. This significant interaction effect remains robust when all covariates are included ($b=-0.51$, $CI_{95}=[-0.94- -0.08]$, $p=0.019$). Conditional effects remain significant when adjusting confidence intervals to control the false discovery rate.

No significant interactions were found for participants' intentions to share the misleading article, whether to express agreement or disagreement, for both debunks and prebunks. The effects remain robust when controlling for the specified control variables outlined in the preregistration and detailed in methods section. Detailed estimates including control variables can be found in Table S-18 and Table S-19 in the Supplementary Material.

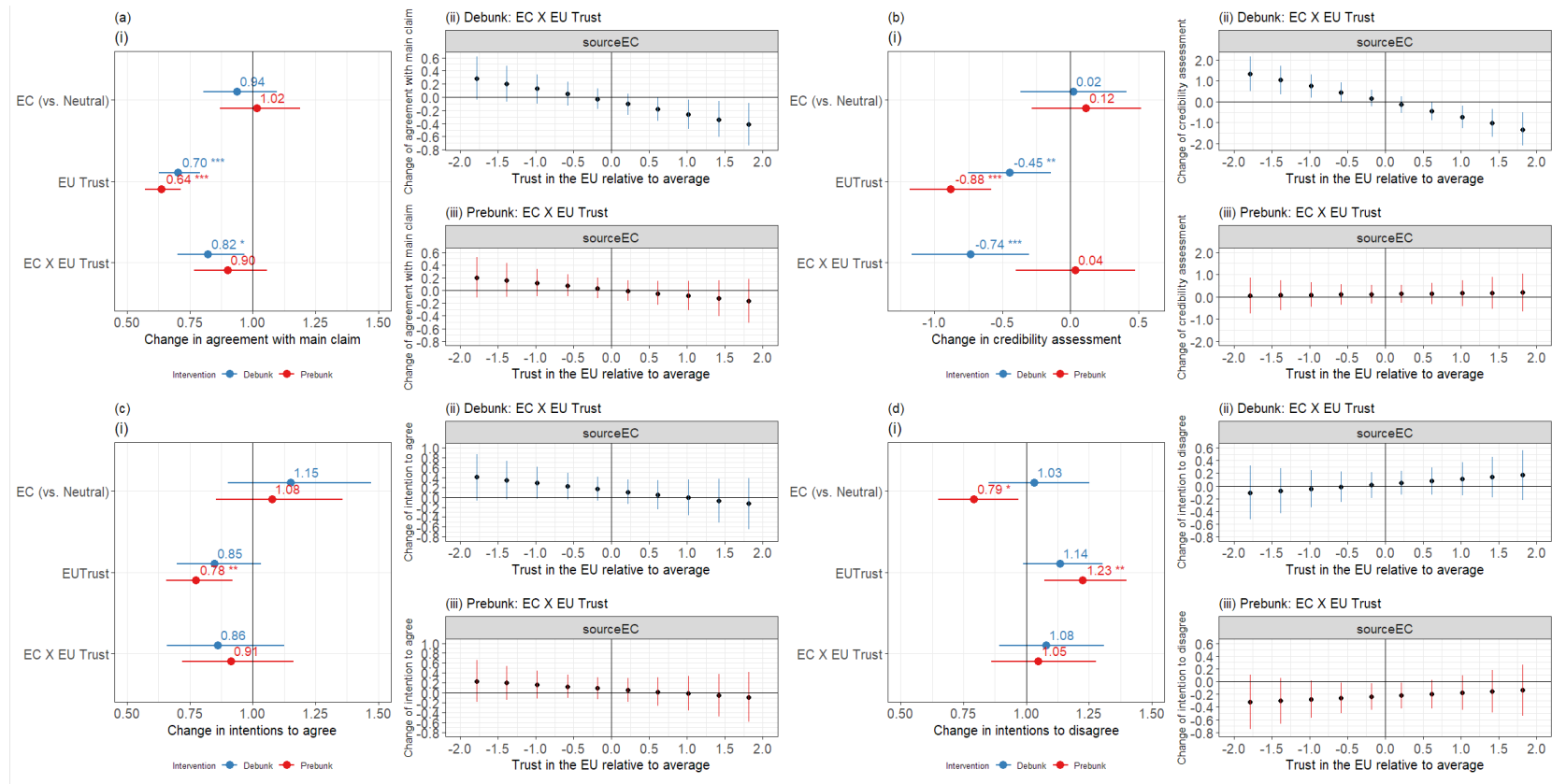


Figure 3. Interactions between trust in the EU and revealing the source of the intervention (i.e. *EC* – European Commission vs. *neutral* – no source) on the main outcome variables for debunks and prebunks, and marginal effects of source reveal conditional on levels of trust in the EU. (i) The y-axis shows the source (*EC* vs. *neutral*), standardized trust in the EU, and the interaction. The x-axis shows the changes in the four main outcome variables; (ii-iii) The y-axis shows the changes in the four outcome variables conditional on trust in the EU ((ii) for debunks (blue) and (iii) for prebunks (red)). The x-axis shows the levels of trust in the EU relative to average trust in standard deviations; (a) shows the effects on agreement with the main claim shown in the misleading article from an ordered logistic regression in (i) as odds ratios and in (ii-iii) as marginal effects; (b) shows the effects on credibility assessments of the misleading article from a linear OLS regression as linear estimates in (i) and in (ii-iii) as marginal effects; (c) (i) shows the effects on intentions to share the misleading article to express agreement with it from a binary logistic regression in (i) as odds ratios and in (ii-iii) as marginal effects; (d) (i) shows the effects on intentions to share the misleading article to express disagreement with it from a binary logistic regression in (i) as odds ratios and in (ii-iii) as marginal effects. Bars represent heteroscedasticity-robust 95% confidence intervals. Significance levels: *** <.001, ** <0.01, * <.05.

Can perceptions of debunks and prebunks explain the effects?

Several mechanisms have been put forward to explain why tailored interventions can be more or less effective than un-tailored ones⁴⁰. To gain insights into these explanations, participants in our experiment evaluated the interventions on dimensions such as perceived relevance, usefulness to improve decision making, authenticity, attention-grabbing nature, and perceived manipulateness. Ratings were recorded on 5-point Likert scales and subsequently dichotomized, with agreement or strong agreement coded as 1 and other responses coded as 0. Binary logistic regression models were then used to estimate the impact of intervention type (prebunk vs. debunk), provided source (EC vs. neutral), and trust in the EU as independent variables. Figure 4 presents the findings.

Our findings reveal several key points. Firstly, prebunks are rated as less relevant, less authentic, and more manipulative compared to debunks. Secondly, revealing that debunks and prebunks come from the European Commission makes participants tend to see them as (slightly) more relevant and authentic. Lastly, individuals with higher levels of trust in the EU are considerably more likely to perceive both interventions as relevant, decision-enhancing, authentic, and attention-grabbing, while being less likely to view them as manipulative.

Further analyses incorporating interactions, similar to those discussed earlier, uncovered two notable cases of significant interaction effects between EC-source and trust in the EU regarding perceptions of prebunks (see Figure S-9 in the Supplementary Material). In the first case, as trust in the EU increases, presenting the EC source becomes increasingly effective in enhancing people's perception of the message's usefulness for making informed decisions. In the second case, as trust in the EU increases, the EC source becomes less likely to be perceived as manipulative.

While these findings regarding perceptions of interventions do not entirely account for the effects observed in the main outcome variables, they do offer some insights. Firstly, the lower effectiveness of prebunks compared to debunks in inducing desired behavioural changes may be attributed to their perceived lack of relevance and authenticity, coupled with a higher perception of manipulative intent. Importantly, these effects persist even when controlling for the source of each intervention. Secondly, although EC-branded interventions are judged as less manipulative and more relevant, decision-enhancing, authentic and attention-grabbing, this does not translate into their higher effectiveness, as explored before. Thirdly, the increased effectiveness of EC debunks, in terms of reducing beliefs and credibility assessments of false or misleading articles among individuals with high trust in the EU, can be partially explained by the finding that as EU trust increases, the inclusion of the EC source enhances the perception of the message as useful for making informed decisions and reduces its perceived manipulative nature.

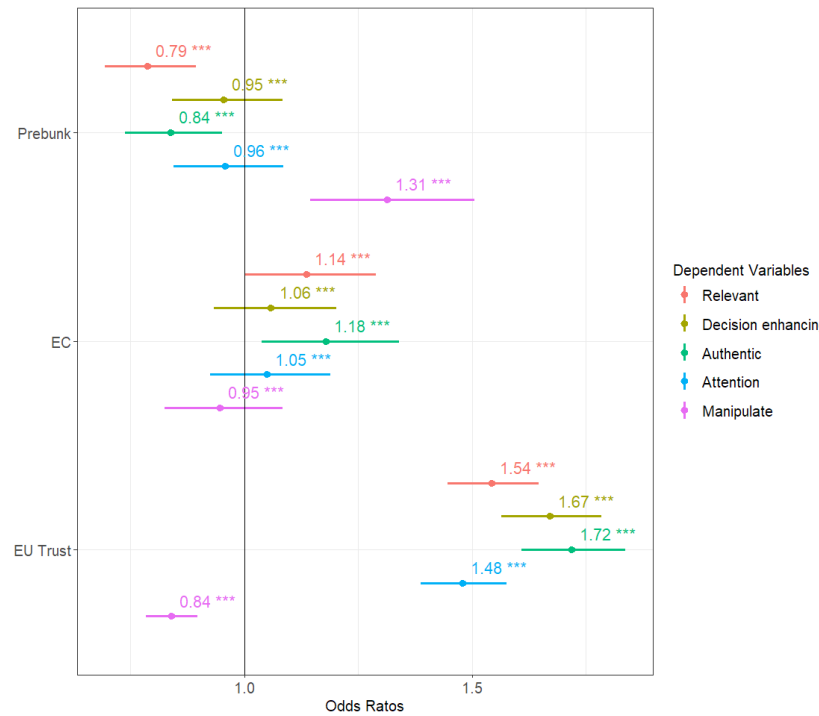


Figure 4. Effects of intervention type, intervention source reveal, and trust in the EU on perceptions of these interventions. Shows the estimates for the effects of the intervention (prebunk vs. debunk), source reveal (EC vs. neutral) and standardized trust in the European Union on perceptions of the interventions with regards to being relevant, decision enhancing, authentic, attention-grabbing, and manipulating. Outcome variables are 1 if participants (strongly) agreed, 0 otherwise. Participants from the control condition are not included, as they saw no intervention they could have rated. EU Trust is demeaned such that a value of 0 on the x-axis corresponds to an average level of trust in the EU. One unit of change of EU trust corresponds to one standard deviation. The models are binary logistic regressions reporting the odds ratios and heteroscedasticity robust 95% confidence intervals. Significance levels: *** <.001, ** <.01, * <.05.

Discussion

First, our results demonstrate that debunking and prebunking interventions effectively address common misinformation claims related to Covid-19 and climate change in Germany, Greece, Ireland, and Poland, expanding upon previous knowledge primarily based on the United States. These interventions persistently influence three out of the four tested outcome variables in the desired direction. Notably, only the EC debunk and the neutral prebunk significantly increase intentions to share a misleading article to express disagreement.

Second, our findings indicate that debunks are slightly more effective than prebunks in combatting misinformation. The two types of intervention do not differ in terms of reducing the perceived credibility of the misleading article claim and increasing people's intention to share the article with others to express their disagreement. However, debunks did reduce beliefs in false claims and intentions to share the misleading articles to endorse them, more so than prebunks. This difference may be attributed to the fact that the employed debunks explicitly address the claims made in the misleading articles while also highlighting commonly used strategies, whereas prebunks solely focus on the latter. Although prebunks therefore have broader applicability, their omission of specifically addressing the false claim that people encounter, and providing a factual substitute may explain their lower effectiveness. Perceptions of the debunking and prebunking interventions shed light on their effects on the main outcome variables,

providing potential explanations. While a causal mediation analysis is not feasible with the current experimental design, the data suggests that prebunks are perceived as more manipulative (with 31.56 % vs. 25.87 % of respondents (strongly) agreeing that the intervention wanted to manipulate them) but less relevant and authentic than debunks, which may account for their reduced effectiveness. These findings indicate that the additional information in debunks, specifically addressing the content of the false claim, serves an important purpose.

Third, the findings show that, on average, revealing the source of the intervention (i.e. the European Commission in our experiment) has virtually no impact on the effectiveness of this intervention. This finding is both reassuring and disappointing for public institutions, policymakers and practitioners. Reassuring, as it means that stamping an intervention with the government sponsor does not hurt the intervention overall. Disappointing, as one may hope that revealing that a governmental body is behind an intervention should increase its positive effect. We find that interventions from the EU are perceived as more relevant and authentic. Therefore, debunking and prebunking interventions remain robust and can be utilized by the EU as a mass-communication tool to counteract misinformation. Whether these findings generalise to institutions like the World Health Organisation (WHO), the Organisation for Economic Co-operation and Development (OECD), or the United Nations (UN) seems plausible but needs to be verified empirically. Our results suggest that revealing the source of the intervention has no impact on its effectiveness. Whether this would also be the case for other sources is unsure.

Lastly, our analyses allow us to disentangle the effects based on people's trust in the source of the intervention. As main effect, we find that trust in the EU is negatively correlated with beliefs in misleading articles, credibility assessments, and intentions to share the misleading article to agree with it, while positively correlated with intentions to share it to disagree. This aligns with existing evidence demonstrating a negative association between institutional trust and susceptibility to conspiracy theories and misinformation^{41–44}. In terms of interaction effects, results show that, as trust in the EU increases, EC debunks are more effective than neutral debunks, for two out of the four outcome variables (i.e. agreement with the false claim and perceived credibility of the false claim). Conversely, neutral debunking surpasses EU debunking among individuals with low levels of trust in the EU. The observed interaction effects cannot be fully explained by the available perception data. Although we do not find significant interactions between EU trust and source reveal for prebunks' effectiveness, we do observe differences in how the intervention is perceived based on individuals' trust in the EU. Specifically, higher trust in the EU is associated with lower perceptions of the intervention as manipulative and higher perceived message usefulness, while lower trust in the EU is linked to lower perceived usefulness of EU interventions.

We recommend investing in trust-building measures to ensure the wide effectiveness of interventions with a revealed source against misinformation across countries. Moreover, it is beneficial to identify population segments with high levels of mistrust and support communication within those populations through direct peer-to-peer communication from trusted sources. Healthcare professionals (HCPs), for instance, are typically perceived as trustworthy providers of health information^{45–47}. Therefore, initiatives to enhance HCPs' skills in debunking vaccination misinformation during patient-HCP interactions could complement the approaches employed in this study⁴⁸.

Knowledge about institutional trust in different segments of the population could be used by institutions to target and tailor prebunks and debunks. For example, selected groups could be addressed with more rigour and with explicitly designed prebunks and debunks – as opposed to a “one-size-fits-all” or

“shotgun” approach where interventions address everyone in the same way. More specifically, in light of evidence that people with low trust in the public institutions are less receptive to the interventions against misinformation, these interventions could either be focused on more receptive segments of the population or be modified to make them more effective for those who are less receptive (or both). These processes correspond to what is known as targeting, namely tailoring in persuasion psychology^{49,50}, and more specifically in health communication^{40,51–55}, communication to reduce climate scepticism⁵⁶, or recently also nudging^{57,58} and debunking⁵⁹.

Targeting and tailoring interventions can enhance their effectiveness by matching specific features with recipient characteristics⁴⁰. Tailored interventions recognize that individuals have different reasons for perceiving, liking, disliking, or reacting to interventions, leading them to prioritize different dimensions of interventions⁴⁰. These interventions can be more relevant, fitting, familiar, fluent, self-efficacy enhancing, authentic, or attention-grabbing. However, tailored messages may also face challenges such as privacy concerns, perceived manipulation, unfair judgments, stereotyping, or repetitiveness^{40,60}. The Facebook-Cambridge Analytica scandal serves as a cautionary example of misusing personal information for targeted campaigns. In 2018, a whistle-blower revealed that Cambridge Analytica used personal information collected without authorisation of data subjects to profile and target them with personalised political advertisement⁶¹. Importantly, people were targeted based on their personality profiles, which were inferred from their likes – a practice which has been shown to work^{62–64}. These targeted campaigns were said to have the objective to influence political preferences and thus elections – in particular the 2016 US Presidential Campaign and the Brexit Referendum. Given the public's negative perception of this event, the use of similar techniques for public policy requires critical assessment, meticulous planning, and transparent implementation.

There are some caveats of our experiment that should be discussed to properly interpret our findings. Firstly, the order in which we measured sharing intentions and beliefs about accuracy may have influenced participants' decision-making⁶⁵. Asking about beliefs beforehand could have prompted participants to consider accuracy, potentially reducing the likelihood of sharing misinformation. However, this should not bias the treatment effects as the order was consistent across all groups. Secondly, the external validity of our experiment is limited as the interventions occurred immediately after exposure to misinformation, without any intermittent stimuli, potentially inflating effect sizes. It is unclear to what extent our findings generalise to more realistic situations of encountering misinformation and to different designs of debunks and prebunks. Thirdly, the slight advantage of debunks over prebunks we observed for some outcome variables could be due to them being implemented at different points in time with respect to encountering the misleading article or due to the difference in content: we cannot unambiguously attribute this behavioural effect to one or the other. However, our findings align with previous findings attesting higher effectiveness to debunks compared to prebunks^{27,28}, while contradicting findings of prebunks being more effective²⁹. Fourthly, participants' self-reported trust in the EU may be influenced by their assigned treatment. Exposure to an EC debunk or prebunk could lead participants to evaluate the EU more favourably later on, potentially due to an experimenter demand effect. Indeed, our analysis shows slightly higher levels of trust among participants in the EC source group, but the difference is not significant ($b = 0.16$, $SE = 0.09$, $p = 0.09$). Additionally, participants in the neutral source (i.e. no revealed source of the intervention) conditions report significantly higher levels of trust in the EU compared to the control (i.e. no intervention) condition ($b = 0.2$, $SE = 0.09$, $p = 0.04$). This does not suggest the presence of an experimenter demand effect. Regression analysis indicates that individuals in the debunking condition

report higher levels of trust in the EU than those in the control condition ($b = 0.26$, $SE = 0.09$, $p = 0.005$), while there is no significant difference for those in the prebunking condition. Further explanation is needed to understand these patterns.

In conclusion, this study highlights the effectiveness of debunking and prebunking interventions in combating misinformation about Covid-19 vaccination and climate change in EU countries. Institutions with the necessary resources, like the European Commission, should prioritize investing in these interventions, potentially targeted or tailored, due to the lack of evidence suggesting the prevalence of unintended effects.

Data availability

The data in support of the findings of this study are available from the Open Science Framework (https://osf.io/7kytz/?view_only=31f586fc34ae42f295038f5db34efcbf, DOI 10.17605/OSF.IO/7KYTZ).

Code availability

The syntax used to analyse the dataset in this study is available from the Open Science Framework (https://osf.io/7kytz/?view_only=31f586fc34ae42f295038f5db34efcbf, DOI 10.17605/OSF.IO/7KYTZ).

References

1. Treen, K. M. d'I., Williams, H. T. P. & O'Neill, S. J. Online misinformation about climate change. *WIREs Clim Change* **11**, (2020).
2. Bruns, H., Dessart, F. J. & Pantazi, M. *Covid-19 misinformation: Preparing for future crises*. (2022).
3. Pummerer, L. *et al.* Conspiracy Theories and Their Societal Effects During the COVID-19 Pandemic. *Social Psychological and Personality Science* **13**, 49–59 (2022).
4. Imhoff, R. & Lamberty, P. A bioweapon or a hoax? The link between distinct conspiracy beliefs about the Coronavirus disease (COVID-19) outbreak and pandemic behavior. *Social Psychological and Personality Science* **11**, 1110–1118 (2020).
5. Loomba, S., Figueiredo, A., Piatek, S. J., Graaf, K. & Larson, H. J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour* **5**, 337–348 (2021).
6. Bursztyn, L., Rao, A., Roth, C. & Yanagizawa-Drott, D. Opinions as Facts. *The Review of Economic Studies* **0**, 1–33 (2022).
7. van der Linden, S. The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. *Personality and Individual Differences* **87**, 171–173 (2015).
8. Lewandowsky, S., Ecker, U. K. H. & Cook, J. Beyond Misinformation: Understanding and Coping with the “Post-Truth” Era. *Journal of Applied Research in Memory and Cognition* **6**, 353–369 (2017).
9. Ecker, U. K. H. *et al.* The psychological drivers of misinformation belief and its resistance to correction. *Nat Rev Psychol* **1**, 13–29 (2022).
10. Lewandowsky, S., Cook, J. & Lombardi, D. Debunking Handbook 2020. *Databrary* (2020) doi:10.17910/b7.1182.
11. Chan, M.-P. S., Jones, C. R., Hall Jamieson, K. & Albarracín, D. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science* **28**, 1531–1546 (2017).

12. Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N. & Cook, J. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* **13**, 106–131 (2012).
13. Cook, J., Lewandowsky, S. & Ecker, U. K. H. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE* **12**, e0175799 (2017).
14. van der Linden, S., Leiserowitz, A., Rosenthal, S. & Maibach, E. Inoculating the Public against Misinformation about Climate Change. *Global Challenges* **1**, 1600008 (2017).
15. Lewandowsky, S. & van der Linden, S. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology* **32**, 348–384 (2021).
16. Traber, C. S., Roozenbeek, J. & van der Linden, S. Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *The ANNALS of the American Academy of Political and Social Science* **700**, 136–151 (2022).
17. van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A. & Lewandowsky, S. Inoculating against misinformation. *Science* **358**, 1141–1142 (2017).
18. Basol, M., Roozenbeek, J. & van der Linden, S. Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition* **3**, 2 (2020).
19. Basol, M. *et al.* Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society* **8**, (2021).
20. Maertens, R., Roozenbeek, J., Basol, M. & van der Linden, S. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology. Applied* **27**, 1–16 (2021).
21. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. *Sci. Adv.* **8**, eabo6254 (2022).
22. Roozenbeek, J. & van der Linden, S. Fake news game confers psychological resistance against online misinformation. *Palgrave Commun* **5**, (2019).
23. Walter, N. & Murphy, S. T. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs* **85**, 423–441 (2018).
24. Walter, N., Brooks, J. J., Saucier, C. J. & Suresh, S. Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media. *Health Communication* **36**, 1776–1784 (2020).
25. Porter, E. & Wood, T. J. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104235118 (2021).
26. Vivion, M. *et al.* Prebunking messaging to inoculate against COVID-19 vaccine misinformation: an effective strategy for public health. *Journal of Communication in Healthcare* **15**, 232–242 (2022).
27. Tay, L. Q., Hurlstone, M. J., Kurz, T. & Ecker, U. K. H. A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British J of Psychology* **113**, 591–607 (2022).
28. Brashier, N. M., Pennycook, G., Berinsky, A. J. & Rand, D. G. Timing matters when correcting fake news. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2020043118 (2021).
29. Jolley, D. & Douglas, K. M. Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *J Appl Soc Psychol* **47**, 459–469 (2017).
30. Petty, R. E. & Cacioppo, J. T. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology* **19**, 123–205 (1986).
31. Pornpitakpan, C. The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *J Appl Social Psychol* **34**, 243–281 (2004).
32. Pennycook, G. & Rand, D. G. The Psychology of Fake News. *Trends in Cognitive Sciences* **25**, 388–402 (2021).

33. Walter, N. & Tukachinsky, R. A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It? *Communication Research* **47**, 155–177 (2020).
34. Amazeen, M. A. & Krishna, A. Processing Vaccine Misinformation: Recall and Effects of Source Type on Claim Accuracy via Perceived Motivations and Credibility. *International Journal of Communication* **17**, 560–582 (2023).
35. Ecker, U. K. H. & Antonio, L. M. Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Mem Cogn* **49**, 631–644 (2021).
36. Guillory, J. J. & Geraci, L. Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition* **2**, 201–209 (2013).
37. European Commission. Tackling coronavirus disinformation.
https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_en (2021).
38. Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J. & Rand, D. G. Quantifying the potential persuasive returns to political microtargeting. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2216261120 (2023).
39. Esarey, J. & Sumner, J. L. Marginal effects in interaction models: Determining and controlling the false positive rate. *Comparative Political Studies* **51**, 1144–1176 (2018).
40. Teeny, J. D., Siev, J. J., Briñol, P. & Petty, R. E. A Review and Conceptual Framework for Understanding Personalized Matching Effects in Persuasion. *J Consum Psychol* **31**, 382–414 (2021).
41. Roozenbeek, J. *et al.* Susceptibility to misinformation about COVID-19 around the world. *R. Soc. open sci.* **7**, 201199 (2020).
42. Eberl, J.-M., Huber, R. A. & Greussing, E. From Populism to the ‘Plandemic’: Why populists believe in COVID-19 conspiracies. *Journal of Elections, Public Opinion and Parties* **31**, 272–284 (2021).
43. Pickles, K. *et al.* COVID-19 Misinformation Trends in Australia: Prospective Longitudinal National Survey. *J Med Internet Res* **23**, e23805 (2021).
44. Šrol, J., Ballová Mikušková, E. & Čvojevová, V. When we are worried, what are we thinking? Anxiety, lack of control, and conspiracy beliefs amidst the COVID-19 pandemic. *Appl Cognit Psychol* **35**, 720–729 (2021).
45. Vraga, E. K. & Bode, L. Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication* **39**, 621–645 (2017).
46. van der Meer, T. G. L. A. & Jin, Y. Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source. *Health Communication* **35**, 560–575 (2020).
47. Durantini, M. R., Albarracín, D., Mitchell, A. L., Earl, A. N. & Gillette, J. C. Conceptualizing the influence of social agents of behavior change: A meta-analysis of the effectiveness of HIV-prevention interventionists for different groups. *Psychological Bulletin* **132**, 212–248 (2006).
48. Baggio, M., Krawczyk, M., Nohlen, H., Pantazi, M. & Proestakis, A. *Applying lessons from behavioural sciences to vaccination acceptance and demand*.
<https://data.europa.eu/doi/10.2760/420194> (2022).
49. Luong, K. T., Garrett, R. K. & Slater, M. D. Promoting Persuasion With Ideologically Tailored Science Messages: A Novel Approach to Research on Emphasis Framing. *Science Communication* **41**, 488–515 (2019).
50. Joyal-Desmarais, K., Rothman, A. J. & Snyder, M. How do we optimize message matching interventions? Identifying matching thresholds, and simultaneously matching to multiple characteristics. *Eur. J. Soc. Psychol.* **50**, 701–720 (2020).
51. Noar, S. M., Benac, C. N. & Harris, M. S. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin* **133**, 673–693 (2007).

52. Schmid, K. L., Rivers, S. E., Latimer, A. E. & Salovey, P. Targeting or tailoring? *Marketing Health Services* **28**, 32–37 (2008).
53. Pink, S. L., Chu, J., Druckman, J. N., Rand, D. G. & Willer, R. Elite party cues increase vaccination intentions among Republicans. *Proceedings of the National Academy of Sciences* **118**, (2021).
54. Mäki, K. O. *et al.* Tailoring interventions to suit self-reported format preference does not decrease vaccine hesitancy. *PLoS ONE* **18**, e0283030 (2023).
55. Habib, G. L. *et al.* The importance of cultural tailoring of communicators and media outlets in an influenza vaccination awareness campaign: a digital randomized trial. *Sci Rep* **13**, 1744 (2023).
56. Dixon, G., Hmielowski, J. & Ma, Y. Improving Climate Change Acceptance Among U.S. Conservatives Through Value-Based Message Targeting. *Science Communication* **39**, 520–534 (2017).
57. Mills, S. Personalized nudging. *Behav. Public Policy* 1–10 (2020) doi:10.1017/bpp.2020.7.
58. Peer, E. *et al.* Nudge Me Right: Personalizing Online Nudges to People’s Decision-Making Styles. *Computers in Human Behavior* **109**, 106347 (2020).
59. Lunz Trujillo, K., Motta, M., Callaghan, T. & Sylvester, S. Correcting Misperceptions about the MMR Vaccine: Using Psychological Risk Factors to Inform Targeted Communication Strategies. *Political Research Quarterly* **74**, 464–478 (2021).
60. Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S. & Herzog, S. M. Public attitudes towards algorithmic personalization and use of personal data online: evidence from Germany, Great Britain, and the United States. *Humanit Soc Sci Commun* **8**, 117 (2021).
61. Cadwalladr, C. & Graham-Harrison, E. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach.
62. Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* **114**, 12714–12719 (2017).
63. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5802–5805 (2013).
64. Walker, C., O’Neill, S. & de-Wit, L. Evidence of Psychological Targeting but not Psychological Tailoring in Political Persuasion Around Brexit. *Exp Results* **1**, e38 (2020).
65. Pennycook, G., Binnendyk, J., Newton, C. & Rand, D. G. A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra: Psychology* **7**, 25293 (2021).
66. Gaziano, C. & McGrath, K. Measuring the concept of credibility. *Journalism Quarterly* **63**, 451–462 (1986).

Supplementary Material

Descriptive analyses

This section outlines the main characteristics of the main dependent variables. Figure S-5 shows that most participants disagreed with the false claim they encountered, but a notable fraction (25.54%) agreed or strongly agreed with it. The majority (25.82%) considered the misinformation completely non-credible across all dimensions. (i.e., as inaccurate, unbelievable, opinionated, and untrustworthy). A significant portion (31.83%) chose not to share or discuss the misleading article, while 26.84% wanted to express (strong) disagreement and 18.36% wanted to signal (strong) agreement.^{***} There were significant associations between all main dependent variables (see Table S-3 for Spearman rank correlations). Notably, agreement with the main claim, credibility assessments and intentions to agree were substantially positively correlated. Intentions to disagree as a reason to share the misleading article were weakly negatively correlated with the other variables.

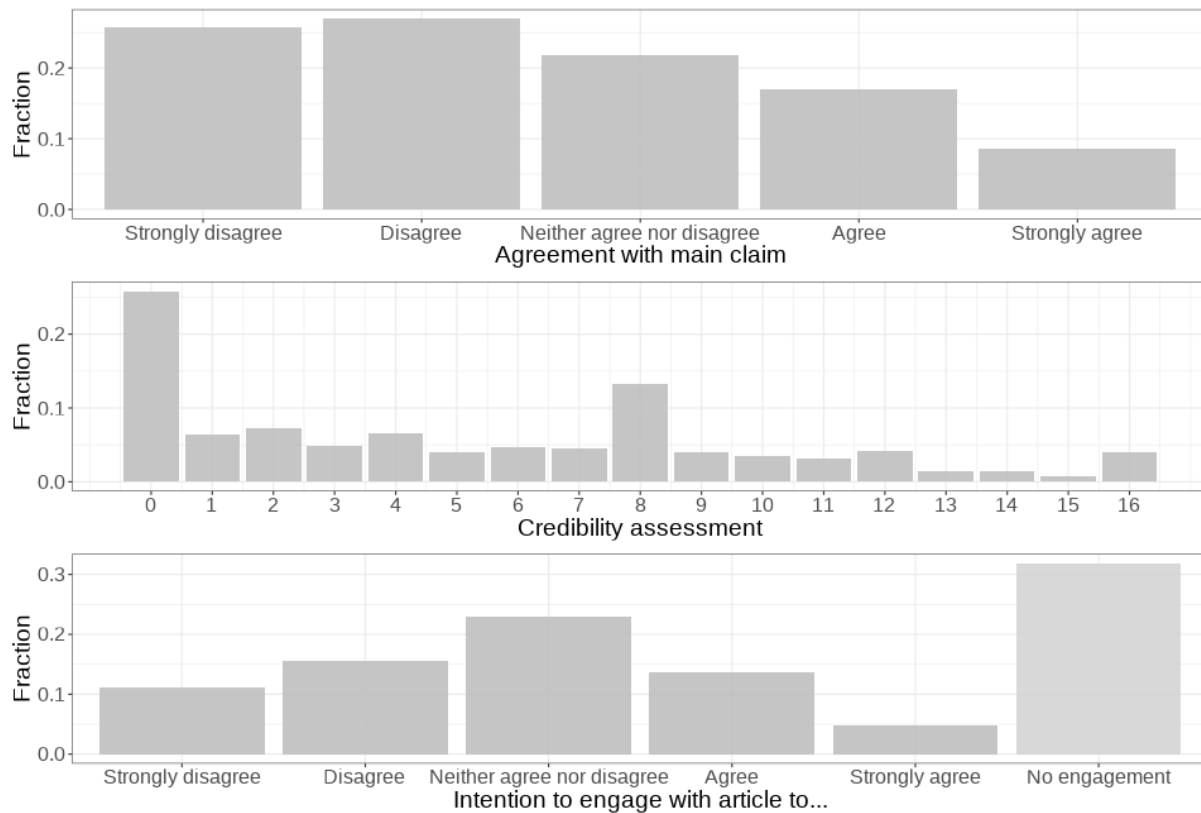


Figure S-5. Distributions of dependent variables. The first panel shows the distribution of ratings of agreement with the main claim of the misleading article participants read. The middle panel depicts the distribution of the added credibility assessments on four dimensions, with response “4” to individual questions meaning the misleading article is considered accurate, believable, factual, and trustworthy respectively, “0” meaning the opposite. A value of 0 here means that the misleading article was rated lowest on all four dimensions, while 16 indicates that the misleading article was rated highest on all four dimensions. The third panel shows the distribution of reasons for intending to share the misleading article for those that indicated that they wanted to

^{***} For the main analyses we do not differentiate sharing intentions according to the intended target group (“people close to you” or “publicly”) or the situation of sharing (“face-to-face” or “online”). The respective distributions of answers are shown in **Figure S-7**.

share it. It ranges from the intention to strongly disagree to the intention to strongly agree. The fraction of participants who did not intend to share the misleading article at all, and consequently did not have to indicate a reason, is also shown.

Table S-2 presents mean values of the main dependent variables, both by treatment and in aggregate, along with the p-values of non-parametric Kruskal-Wallis tests. The distributions of these variables differed by treatment. Each intervention led to a decrease in agreement with the main claim, assessment of the misleading article, and the cumulative credibility ratings. Likewise, the proportions of participants intending to share or discuss the misleading article to (strongly) agree with its main claim were lower in the treatment groups. Conversely, the fractions of individuals planning to share or talk about the misleading article to (strongly) disagree with it were slightly higher among the groups receiving an intervention.

Figure S-6 displays the distribution of trust in the EU, which is the main moderator of interest for the main analyses. On average, the trust level is 5.46 (SD=2.5). Clearly, the variable is not normally distributed (Kolmogorov-Smirnov test: $D = 0.87$, $p < 0.001$) with many observations (around 10%) corresponding to minimal trust in the EU. Trust levels differed significantly among the four countries ($p < 0.000$ in a Kruskal-Wallis test), with Greece having the lowest values and Ireland having the highest (see Figure S-8).

Table S-2. Mean values of dependent variables, by treatment. Shows the means for the respective dependent variables by treatments and for the overall sample. Standard deviations are shown in brackets. The last row contains p values of a non-parametric Kruskal-Wallis test with null hypothesis being that mean ranks are the same in all the treatments.

Treatment	Agreement with the Main Claim	Credibility Assessment	Intention to agree	Intention to disagree
Control	2.9 (1.31)	6.15 (4.85)	0.24 (0.43)	0.24 (0.43)
EC Debunk	2.38 (1.21)	4.9 (4.67)	0.16 (0.37)	0.29 (0.45)
EC Prebunk	2.54 (1.29)	5.01 (4.73)	0.19 (0.4)	0.25 (0.43)
Neutral Debunk	2.39 (1.2)	4.84 (4.51)	0.14 (0.35)	0.28 (0.45)
Neutral Prebunk	2.55 (1.27)	4.95 (4.63)	0.18 (0.38)	0.29 (0.45)
Total	2.56 (1.27)	5.18 (4.71)	0.18 (0.39)	0.27 (0.44)
Range	(s. dis)1-5(s.agree)	(incred.)0-16(cred.)	(no)0-1(yes)	(no)0-1(yes)
Kwallis all treats: p	<.001	<.001	<.001	0.026

Table S-3. Spearman rank correlations between dependent variables.

	Agreement with the Main Claim	Credibility Assessment	Intention to agree	Intention to disagree
Agreement with the Main Claim	1			
Credibility Assessment	.6652	1		
Intention to agree	.4508	.5308	1	
Intention to disagree	-.2759	-.2486	-.2872	1

All correlations were significant at $p < .001$, also after Bonferroni correction for multiple tests.

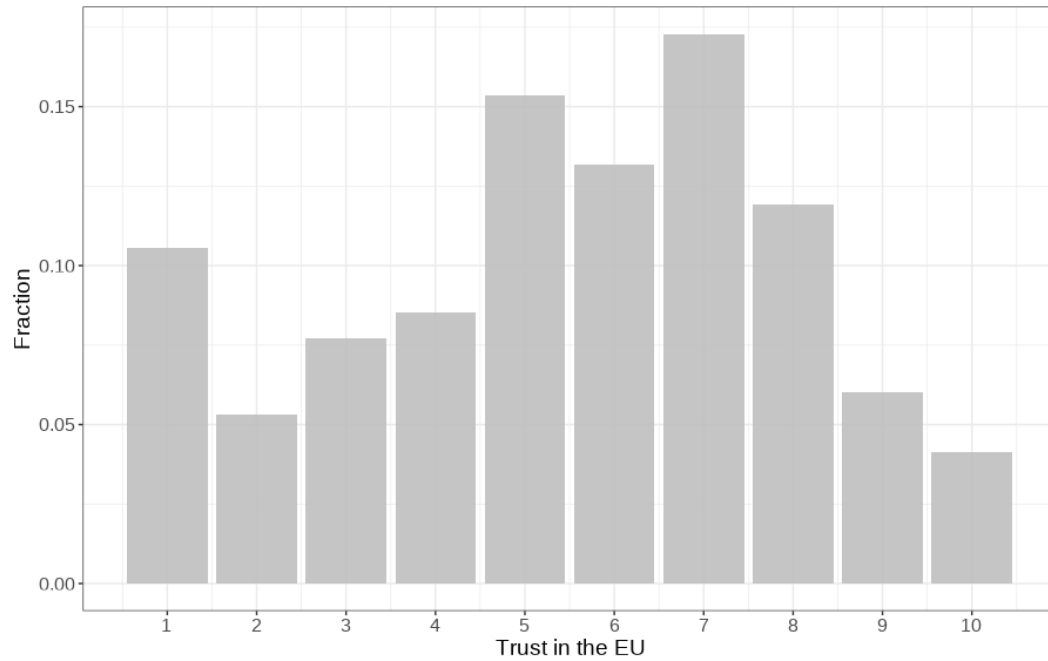


Figure S-6. Distribution of EU trust. Shows the fractions of respondents with the indicated levels of trust in the European Union, ranging from 1 (lowest level of trust) to 10 (highest level of trust). Data on EU trust is missing for 1.4% of the sample.

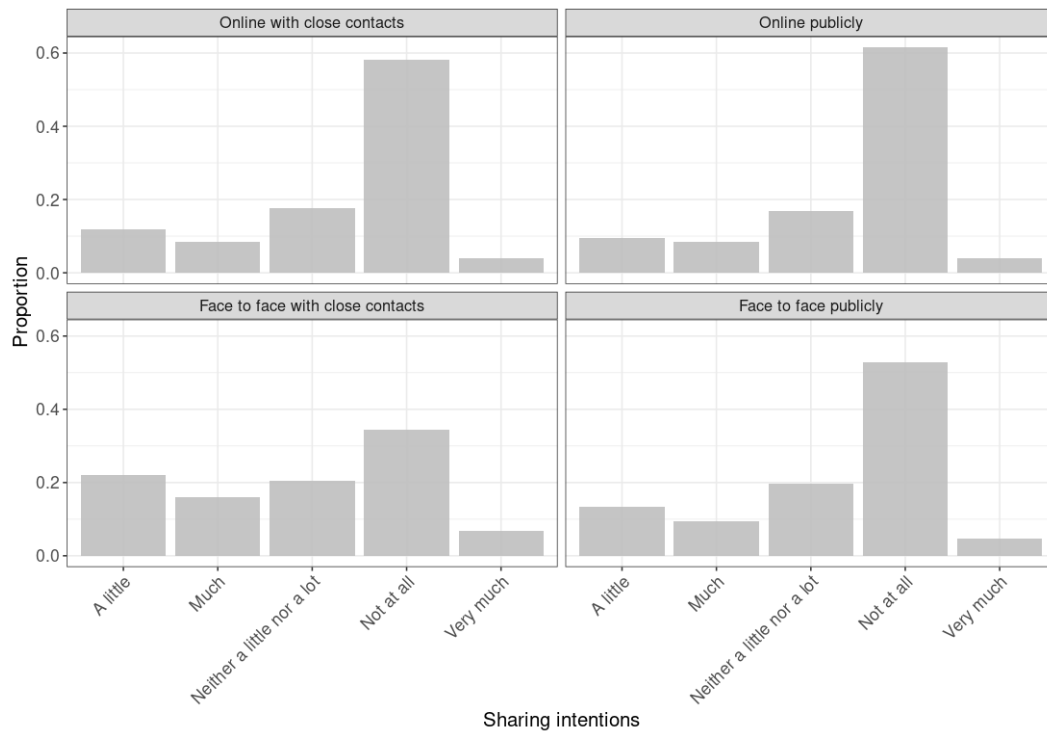


Figure S-7. Distribution of specific sharing intentions. Shows proportions with respect to the people that indicate that they would share the misleading article.

Table S-4. Trust in the European Union by country. Shows the means of trust in the European Union by country and for the overall sample. Standard deviations are shown in brackets. The last row contains p values of a non-parametric Kruskal-Wallis test with null hypothesis being that mean ranks are the same in all the treatments. There are 73 cases where respondents did not indicate trust. A Chi-squared test indicates that missing variables are random with respect to countries ($X^2(4)=7.35$, $p=0.119$).

Country	Trust in the EU
Germany	5.29 (2.43)
Greece	4.96 (2.49)
Ireland	5.99 (2.21)
Poland	5.62 (2.72)
Total	5.46 (2.5)
Range	1-10
Kwallis all treats: p	<.001

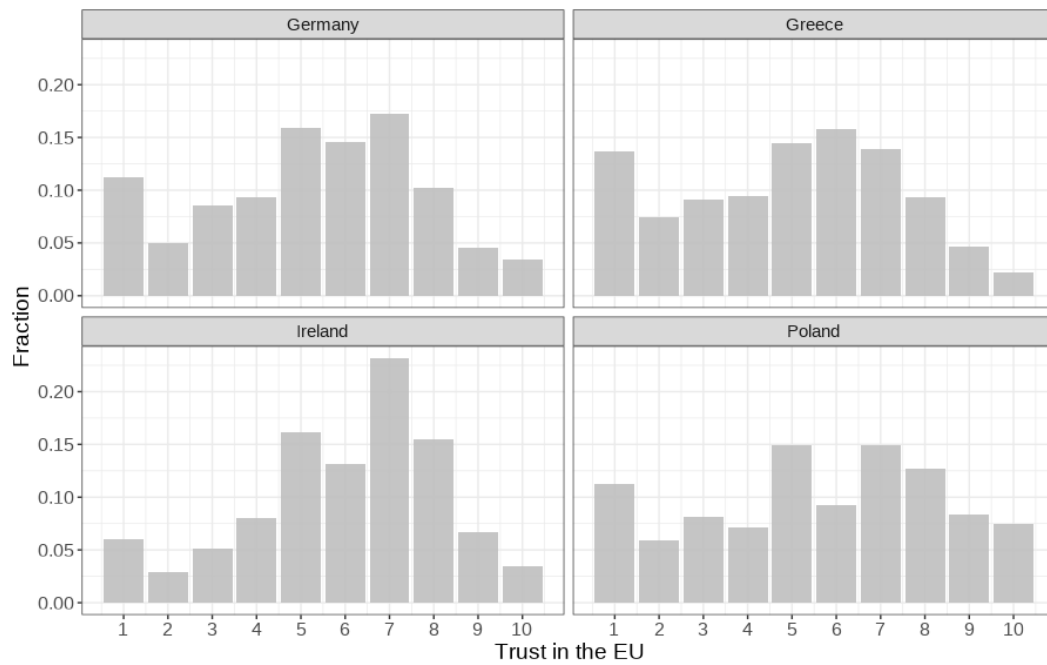


Figure S-8. Distributions of trust in the European Union by country. Shows the fractions of respondents with the indicated levels of trust in the European Union, ranging from 1 (lowest level of trust) to 10 (highest level of trust), for the four different countries.

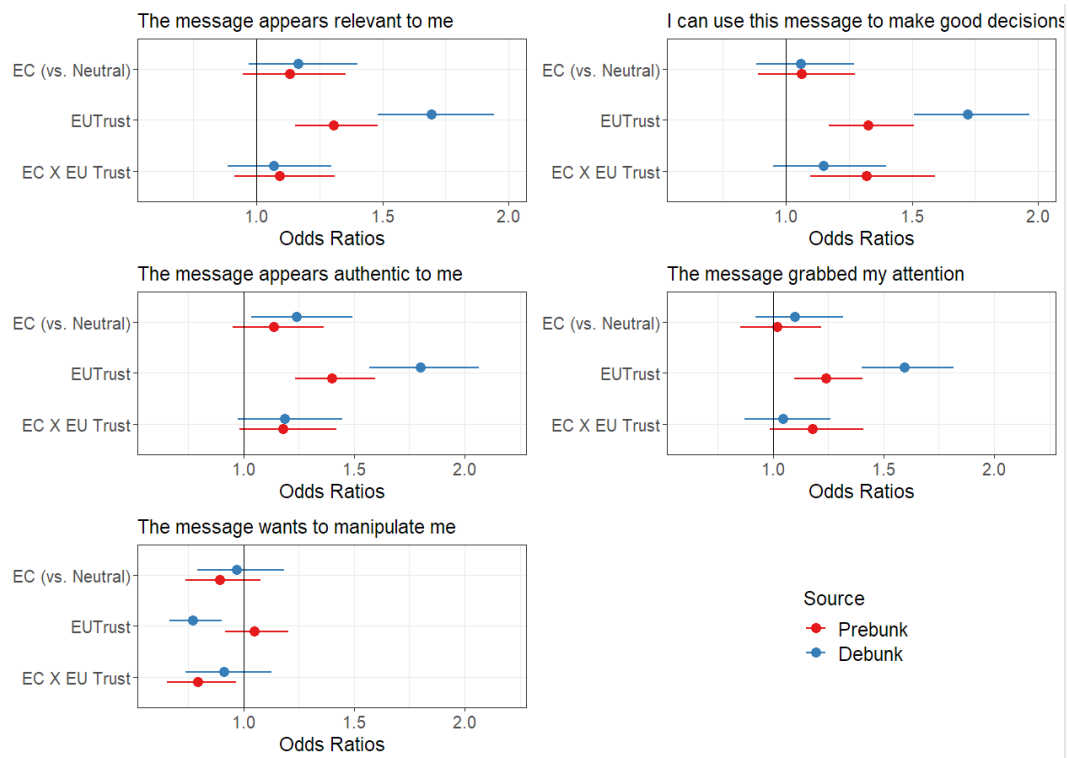


Figure S-9. Interactions between EU trust and source branding on perceptions of debunks and prebunks. Shows the estimates for the effects of EC source for people with average levels of trust in the EU, the correlation between standardized EU trust and the outcome for people in the neutral source treatment, and interaction of the source information and trust in the European

Union, both for debunks (blue) and prebunks (red) on perceptions of the interventions with regards to being relevant, decision enhancing, authentic, attention-grabbing, and manipulating. Outcome variables are 1 if participants (strongly) agreed, 0 otherwise. Participants from the control condition are not included, as they saw no intervention they could have rated. EU Trust is demeaned such that a value of 0 on the x-axis corresponds to an average level of trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. The models are binary logistic regressions reporting the odds ratios and heteroscedasticity robust 95%. Significance levels: *** <.001, ** <0.01, * <.05.

Comparing misinformation on Covid-19 and climate change

The participants were divided into two groups, with half reading articles containing misleading claims about Covid-19 and the other half reading articles with misleading claims about climate change. To examine if the effectiveness of our interventions varies depending on the topic, separate regressions were conducted for Covid-19 and climate change. The relevant tables and plots are provided below: Table S-9 presents the main effects with interaction between intervention and topic, while Table S-10 and Table S-11 show the focal interaction effects interacted with topic. Figure S-10 displays the main effects by topic, and Figure S-11 - Figure S-14 depict the interaction effects by topic. Additionally, Figure S-15 - Figure S-18 illustrate the effects of branding conditional on levels of EU trust for each topic.

Since no specific hypotheses were pre-registered regarding the moderating effect of the topic, these findings should be considered exploratory. The main effects of interventions on agreement with the main claim were nearly identical for Covid-19 and climate change. However, the effects for the remaining three dependent variables (credibility assessment, intention to agree, intention to disagree) were generally weaker for climate change compared to Covid-19. Both debunks and prebunks effectively reduced credibility assessments for both topics, but the effects were less pronounced for climate change (though not statistically significantly so). The same pattern occurs for behavioural intentions. Interestingly, all interventions effectively reduced intentions to agree with the main claim for Covid-19, while only the neutral debunk was effective for climate change. Regarding intentions to disagree, only the debunks effectively increased them for Covid-19, but none of the interventions had a similar effect for climate change. Although prebunks and debunks appeared to be more effective in addressing Covid-19 misinformation overall, none of the interaction effects reached significance (see Table S-9 and Figure S-10).

To examine if our main interactions of interest between source branding and trust in the EU are sensitive to the topic of misinformation, we visually inspected forest plots showing effects of providing source information, EU trust and their interaction separately for interventions and the topic (Figure S-11 - Figure S-14), and conditional effect plots (Figure S-15 - Figure S-18). We also estimated models with a three-way-interaction between source, EU trust and topic (see Table S-10 and Table S-11). Despite some visual differences, none of the interactions were statistically significant. Notably, the significant interaction of source branding for debunks mentioned above is insignificant for both climate change and Covid-19 (albeit slightly stronger for Covid-19). By pooling the data for both topics, the narrower confidence intervals allow for more precise estimates, therefore leading to a significant interaction when aggregating over the topic. Judging by the conditional effect plots, the more pronounced effect for EC-branded debunks for people with high trust in the EU appears to occur only for the Covid-19 topic. The opposite is the case for the interaction of source-branding of the debunk with EU trust on credibility assessments. The “backfire effect” of EC branding for people with low trust in the EU and the higher effectiveness for high-trust individuals appears to occur mainly for climate change.

Regarding intentions to share the misleading article or talk about it, there are no major differences with respect to the topic. Notably, except for the effects of debunks on intentions to disagree (Figure S-14), the interactions go into opposite directions both for Covid-19 and climate change (Figure S-11 - Figure S-13). The interaction is qualitatively counter-intuitive in some cases, as can be seen from the slightly positive slopes of the conditional effect plots (see Figure S-15 and Figure S-16 for climate change, and Figure S-17 and Figure S-18 for Covid-19).

Interestingly, there are instances where the interaction between topic and EU trust is significant (see Table S-10 and Table S-11). However, there is no consistent pattern regarding the significance and direction of these interactions. Two interactions are positive, indicating that the association between trust in the EU and the outcome variable is stronger for climate change compared to Covid-19. These cases include prebunk effects on agreement with the main claim ($OR=1.36$, $CI_{95}=[1.08-1.71]$, $p=0.009$) and debunk effects on intentions to agree ($OR=1.58$, $CI_{95}=[1.11-2.25]$, $p=0.023$). In two cases, the interaction is negative, suggesting that the correlation between trust in the EU and the outcome variable is more pronounced for misleading Covid-19 than for misleading climate change articles. These cases include prebunk effects on agreement with the main claim ($OR=0.74$, $CI_{95}=[0.12-1.35]$, $p=0.019$) and prebunk effects on intentions to disagree ($OR=0.7$, $CI_{95}=[0.53-0.93]$, $p=0.01$).

Table S-5. Results from models for the four main outcome variables. Shows the estimates for the intervention effects with the Control condition as the baseline. *Model 1 is an ordered logistic regression reporting the odds ratios. Model 2 reports linear estimates from an OLS model. Models 3 and 4 report odds ratios from a binary logistic regression. For all models, heteroscedasticity-robust confidence intervals and p-values are provided. Effects for EU Trust represent the change of one standard deviation of EU Trust on the respective DV.*

<i>Predictors</i>	Agreement with the Main Claim			Credibility Assessment			Intention to agree			Intention to disagree		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
Intercept				6.10	5.82 – 6.39	<0.001	0.30	0.26 – 0.35	<0.001	0.32	0.28 – 0.37	<0.001
Neutral Debunk	0.51	0.44 – 0.60	<0.001	-1.20	-1.60 – -0.81	<0.001	0.54	0.43 – 0.68	<0.001	1.19	0.98 – 1.45	0.081
EC Debunk	0.48	0.41 – 0.56	<0.001	-1.22	-1.62 – -0.83	<0.001	0.64	0.51 – 0.79	<0.001	1.24	1.02 – 1.51	0.028
Neutral Prebunk	0.59	0.51 – 0.69	<0.001	-1.22	-1.62 – -0.82	<0.001	0.70	0.57 – 0.87	0.001	1.26	1.04 – 1.53	0.021
EC Prebunk	0.61	0.52 – 0.71	<0.001	-1.10	-1.50 – -0.70	<0.001	0.77	0.62 – 0.95	0.014	1.00	0.82 – 1.23	0.962
EUTrust	0.62	0.58 – 0.66	<0.001	-0.92	-1.05 – -0.78	<0.001	0.75	0.70 – 0.80	<0.001	1.23	1.15 – 1.31	<0.001
Observations		5155			5155			5155			5155	
R2 Nagelkerke		0.129			0.049 / 0.048			0.022			0.01	

Intercepts for ordered logit model are: Strongly disagree | disagree: 0.2, (0.18 – 0.23), p<0.001; Disagree | Neither agree nor disagree: 0.7, (0.63 – 0.79), p<0.001; Neither agree nor disagree | Agree: 1.93, (1.72 – 2.17), p<0.001; Agree | Strongly agree: 7.52, (6.57 – 8.61), p<0.001.

Table S-6. Effects of debunks, prebunks, and EC-branding on the main outcome variables. Shows the estimates for the effects of the debunk and prebunk vs. the control, for providing the EC as the source with the neutral-source condition as the baseline, and for standardized trust in the EU. Model 1 is an ordered logistic regression reporting the odds ratios. Model 2 reports linear estimates from an OLS model. Models 3 and 4 report odds ratios from a binary logistic regression. For all models, heteroscedasticity-robust confidence intervals and p-values are provided. Effects for EU Trust represent the change of one standard deviation of EU Trust on the respective DV.

Predictors	Agreement with the Main Claim			Credibility Assessment			Intention to agree			Intention to disagree		
	Odds Ratios	CI	p	Estimates	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p
Intercept				6.10	5.82 – 6.39	<0.001	0.30	0.26 – 0.35	<0.001	0.32	0.28 – 0.37	<0.001
Debunk	0.50	0.44 – 0.58	<0.001	-1.24	-1.61 – -0.87	<0.001	0.55	0.45 – 0.68	<0.001	1.27	1.06 – 1.52	0.010
Prebunk	0.61	0.53 – 0.70	<0.001	-1.19	-1.56 – -0.81	<0.001	0.69	0.57 – 0.84	<0.001	1.18	0.98 – 1.42	0.083
EC (vs. Neutral)	0.98	0.87 – 1.09	0.659	0.05	-0.23 – 0.33	0.730	1.13	0.96 – 1.33	0.148	0.92	0.80 – 1.05	0.219
EU Trust	0.62	0.59 – 0.65	<0.001	-0.91	-1.05 – -0.77	<0.001	0.75	0.70 – 0.80	<0.001	1.22	1.15 – 1.30	<0.001
Observations		5155			5155			5155			5155	
R2 Nagelkerke		0.128			0.049 / 0.049			0.022			0.009	

Intercepts for ordered logit model for Agreement as DV are: Strongly disagree | disagree: 0.20, (0.18 – 0.23), p<0.001; Disagree | Neither agree nor disagree: 0.70, (0.63 – 0.79), p<0.001; Neither agree nor disagree | Agree: 1.93, (1.73 – 2.16), p<0.001; Agree | Strongly agree: 7.52, (6.59 – 8.59), p<0.001.

Table S-7. Effects of EC-branded vs. neutrally branded interventions and interaction with EU trust on beliefs and credibility ratings. Shows the estimates for the effects of providing the EC as the source with the neutral-source condition as the baseline. *Models 1 and 2 are ordered logistic regression reporting the odds ratios. Models 3 and 4 report linear estimates from OLS models. For all models, heteroscedasticity-robust confidence intervals and p-values are provided.*

Intervention	Agreement with the Main Claim						Credibility Assessment					
	Debunk			Prebunk			Debunk			Prebunk		
Predictors	Odds Ratios	CI	p	Odds Ratios	CI	p	Estimates	CI	p	Estimates	CI	p
Intercept							4.86	4.58 – 5.13	<0.001	4.88	4.60 – 5.17	<0.001
EC vs. Neutral Intervention	0.94	0.80 – 1.10	0.422	1.02	0.87 – 1.19	0.833	0.02	-0.37 – 0.41	0.904	0.12	-0.28 – 0.52	0.565
EUTrust	0.70	0.62 – 0.80	<0.001	0.64	0.57 – 0.72	<0.001	-0.45	-0.75 – -0.14	0.004	-0.88	-1.18 – -0.58	<0.001
EC Intervention X EUTrust	0.82	0.69 – 0.98	0.033	0.90	0.76 – 1.07	0.238	-0.74	-1.17 – -0.30	0.001	0.04	-0.40 – 0.48	0.866
Observations	2066			2012			2066			2012		
R2 Nagelkerke	0.115			0.105			0.038 / 0.036			0.035 / 0.033		

Intercepts for ordered logit model of Debunks are: Strongly disagree | disagree: 0.37, (0.32 – 0.41), p<0.001; Disagree | Neither agree nor disagree: 1.42, (1.27 – 1.60), p<0.001; Neither agree nor disagree | Agree: 4.13, (3.60 – 4.72), p<0.001; Agree | Strongly agree: 15.56, (12.80 – 18.92), p<0.001.

Intercepts for ordered logit models of Prebunks are: Strongly disagree | disagree: 0.36, (0.32 – 0.41), p<0.001; Disagree | Neither agree nor disagree: 1.15, (1.02 – 1.29), p=0.023; Neither agree nor disagree | Agree: 3.11, (2.72 – 3.55), p<0.001; Agree | Strongly agree: 12.54, (10.44 – 15.06), p<0.001.

Table S-8. Effects of EC-branded vs. neutrally branded interventions and interaction with EU trust on intentions to endorse and criticise. Shows the estimates for the effects of providing the EC as the source with the neutral-source condition as the baseline. All models report odds ratios from binary logistic regressions. For all models, heteroscedasticity-robust confidence intervals and p-values are provided.

Intervention	Intention to agree						Intention to disagree					
	Debunk			Prebunk			Debunk			Prebunk		
Predictors	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p	Ratios	CI	p
Intercept	0.17	0.14 – 0.20	<0.001	0.22	0.18 – 0.25	<0.001	0.39	0.34 – 0.45	<0.001	0.41	0.35 – 0.47	<0.001
EC vs. Neutral Intervention	1.15	0.90 – 1.47	0.254	1.08	0.86 – 1.36	0.527	1.03	0.85 – 1.25	0.750	0.79	0.65 – 0.97	0.023
EUTrust	0.85	0.71 – 1.01	0.100	0.78	0.66 – 0.91	0.003	1.14	0.99 – 1.31	0.072	1.23	1.07 – 1.41	0.003
EC Intervention X EUTrust	0.86	0.68 – 1.09	0.276	0.91	0.73 – 1.14	0.468	1.08	0.89 – 1.31	0.429	1.05	0.86 – 1.28	0.631
Observations	2066			2012			2066			2012		
R2 Nagelkerke	0.011			0.016			0.006			0.012		

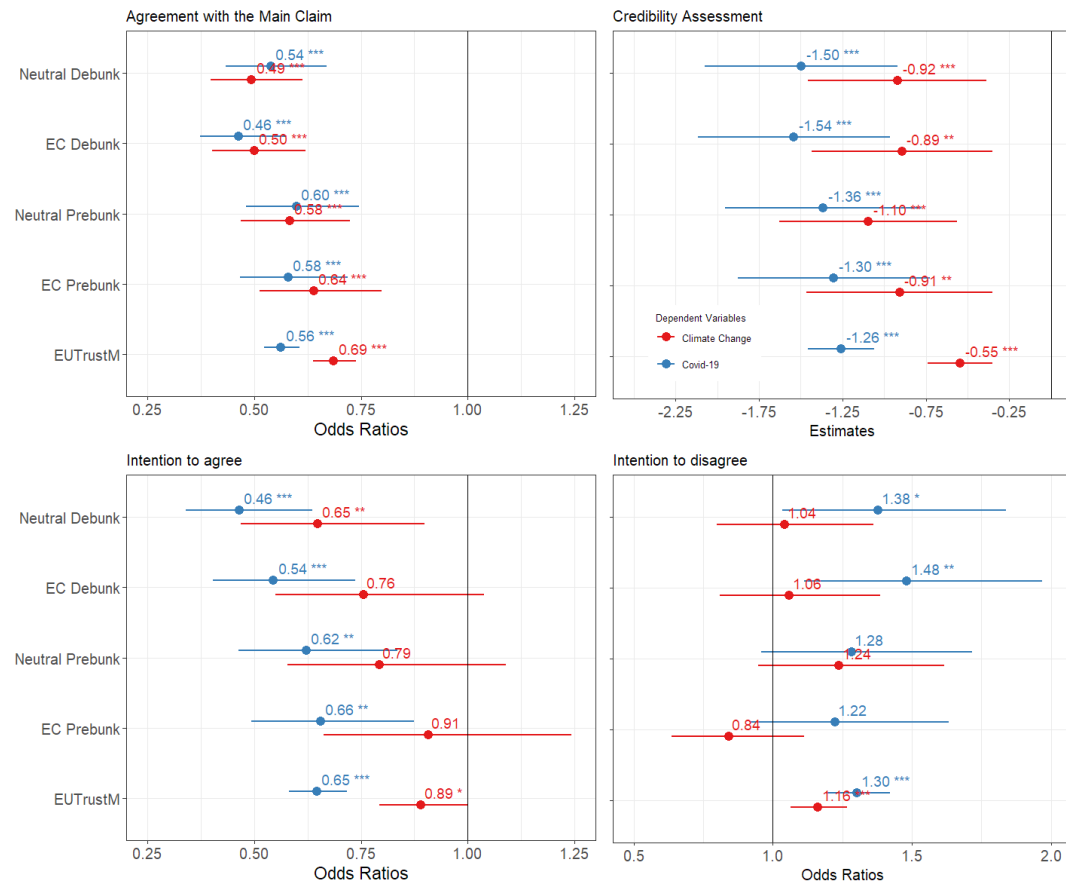


Figure S-10. Treatment effects on main outcome variables by misleading article topic (climate change in red and Covid-19 in blue). The y-axis shows the interventions (with control as the reference condition), and standardized trust in the EU. The x-axis shows the changes in the four main outcome variables. (a) shows the effects on agreement with the main claim shown in the misleading article from an ordered logistic regression as odds ratios; (b) shows the effects on credibility assessments of the misleading article from a linear OLS regression as linear estimates; (c) shows the effects on intentions to agree with the misleading article from a binary logistic regression as odds ratios; (d) shows the effects on intentions to disagree with the misleading article from a binary logistic regression as odds ratios. Effects for climate change misinformation are shown in red, for Covid-19 in blue. Bars represent heteroscedasticity-robust 95% confidence intervals. Significance levels: *** <.001, ** <0.01, * <.05.

Table S-9. Results from models for the four main outcome variables showing interactions of misleading article topic and intervention. Shows the estimates for the intervention effects with the Control condition as the baseline, interacted with the topic of the misinformation (climate change vs. Covid-19, the latter being the baseline). *Model 1 is an ordered logistic regression reporting the odds ratios. Model 2 reports linear estimates from an OLS model. Models 3 and 4 report odds ratios from a binary logistic regression. For all models, heteroscedasticity-robust confidence intervals and p-values are provided.*

Predictors	Agreement with main claim			Credibility Assessment			Intention to agree			Intention to disagree		
	Odds Ratios	CI	p	Estimates	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p
Neutral Debunk	0.53	0.42 – 0.66	<0.001	-1.55	-2.12 – -0.97	<0.001	0.46	0.34 – 0.63	<0.001	1.39	1.04 – 1.85	0.025
EC Debunk	0.46	0.37 – 0.57	<0.001	-1.56	-2.14 – -0.99	<0.001	0.55	0.40 – 0.73	<0.001	1.48	1.12 – 1.97	0.006
Neutral Prebunk	0.60	0.48 – 0.75	<0.001	-1.36	-1.95 – -0.76	<0.001	0.63	0.47 – 0.84	0.002	1.28	0.96 – 1.71	0.095
EC Prebunk	0.57	0.46 – 0.72	<0.001	-1.32	-1.90 – -0.74	<0.001	0.66	0.49 – 0.87	0.003	1.23	0.92 – 1.64	0.163
Climate Change (vs. Covid-19)	0.99	0.80 – 1.23	0.936	-0.40	-0.96 – 0.16	0.166	0.63	0.47 – 0.84	0.002	1.48	1.12 – 1.96	0.006
EU Trust	0.62	0.59 – 0.66	<0.001	-0.91	-1.05 – -0.77	<0.001	0.75	0.70 – 0.80	<0.001	1.23	1.15 – 1.31	<0.001
Neutral Debunk X Climate Change	0.95	0.70 – 1.29	0.753	0.69	0.10 – 1.47	0.088	1.44	0.92 – 2.26	0.114	0.74	0.50 – 1.10	0.140
EC Debunk X Climate Change	1.10	0.81 – 1.48	0.547	0.69	0.10 – 1.48	0.089	1.39	0.90 – 2.15	0.139	0.71	0.48 – 1.05	0.088
Neutral Prebunk X Climate Change	0.98	0.71 – 1.34	0.885	0.27	0.53 – 1.07	0.504	1.26	0.82 – 1.94	0.296	0.97	0.65 – 1.43	0.862
EC Prebunk X Climate Change	1.12	0.82 – 1.54	0.472	0.44	0.36 – 1.25	0.279	1.40	0.92 – 2.14	0.120	0.68	0.46 – 1.02	0.061
(Intercept)				6.30	5.89 – 6.72	<0.001	0.38	0.31 – 0.46	<0.001	0.26	0.21 – 0.32	<0.001
Observations	5155			5155			5155			5155		
R ² Nagelkerke	0.129			0.050 / 0.048			0.025			0.012		

Intercepts for ordered logit model of agreement with the main claim are: Strongly disagree | disagree: 0.20, (0.17 – 0.24), p<0.001; Disagree | Neither agree nor disagree: 0.7, (0.60 – 0.82), p<0.001; Neither agree nor disagree | Agree: 1.92, (1.64 – 2.26), p<0.001; Agree | Strongly agree: 7.49, (6.26 – 8.97), p<0.001.

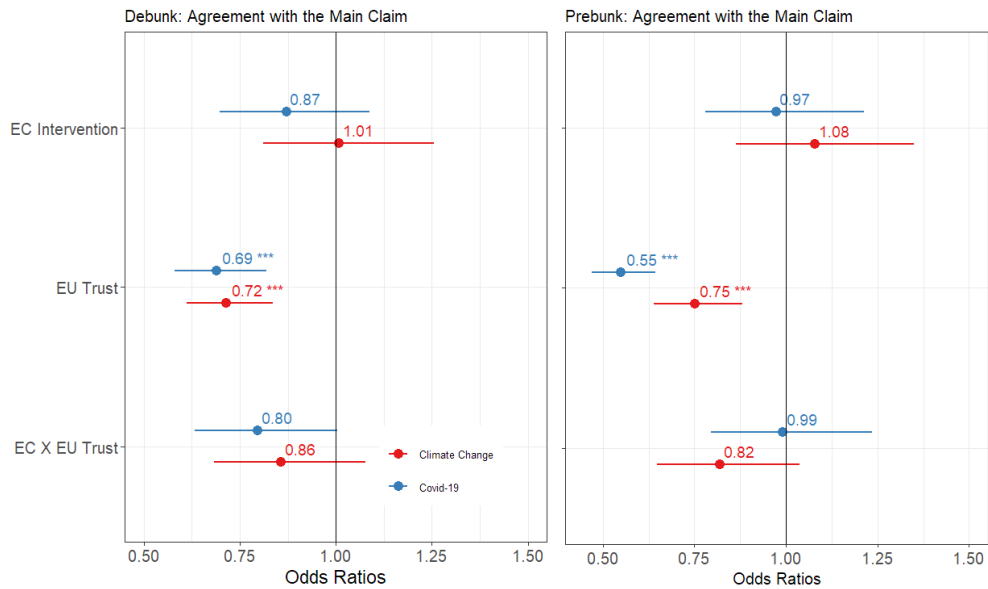


Figure S-11. Treatment effects for average levels of trust in the European Union and interaction between source treatment and trust in the European Union on agreement with the main claim, by misleading article topic. Shows the estimates for the interaction of the source information and trust in the European Union, both for debunks (left) and prebunks (right), for climate change in red and Covid-19 in blue. EU Trust is demeaned such that a value of 0 on the x-axis corresponds to an average level of trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. The model is an ordered logistic regression reporting the odds ratios and heteroscedasticity robust 95%. Significance levels: *** <.001, ** <.01, * <.05.

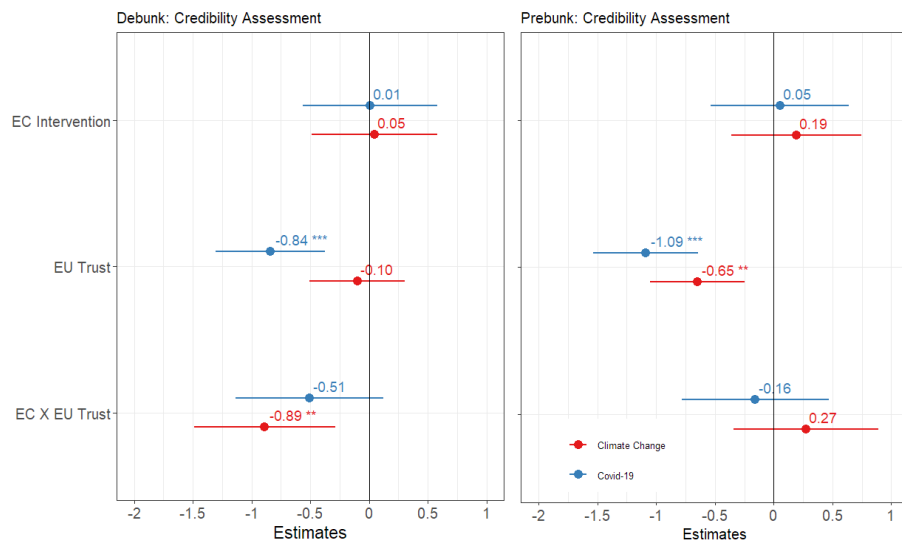


Figure S-12. Treatment effects for average levels of trust in the European Union and interaction between source treatment and trust in the European Union on credibility assessment, by misleading article topic. Shows the estimates for the interaction of the source information and trust in the European Union, both for debunks (left) and prebunks (right), for climate change in red and Covid-19 in blue. EU Trust is demeaned such that a value of 0 on the x-axis corresponds to an average level of trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. The model is a linear OLS regression reporting linear estimates and heteroscedasticity robust 95%. Significance levels: *** <.001, ** <.01, * <.05.

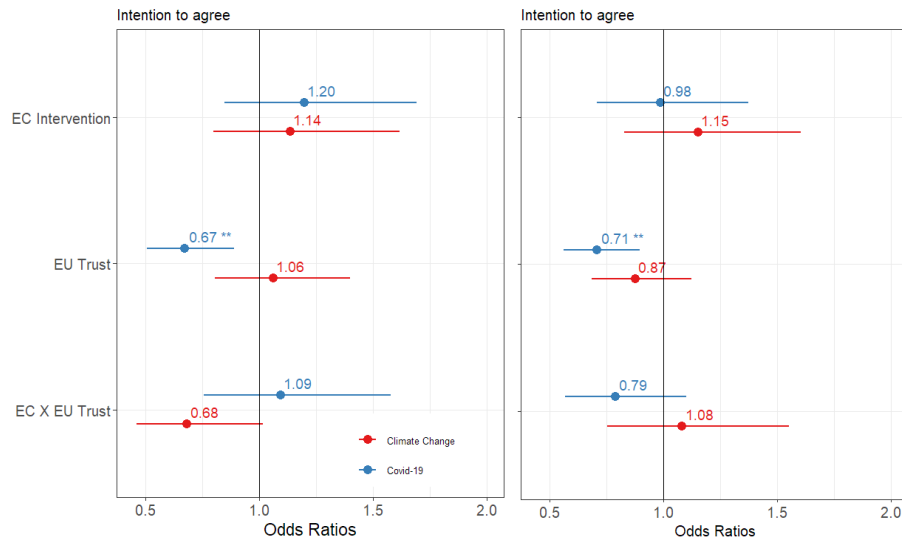


Figure S-13. Treatment effects for average levels of trust in the European Union and interaction between source treatment and trust in the European Union on likelihood to express agreement with misleading article, by article topic. Shows the estimates for the interaction of the source information and trust in the European Union, both for debunks (left) and prebunks (right), for climate change in red and Covid-19 in blue. EU Trust is demeaned such that a value of 0 on the x-axis corresponds to an average level of trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. The model is a binary logistic regression reporting the odds ratios and heteroscedasticity robust 95%. Significance levels: *** <.001, ** <0.01, * <.05.

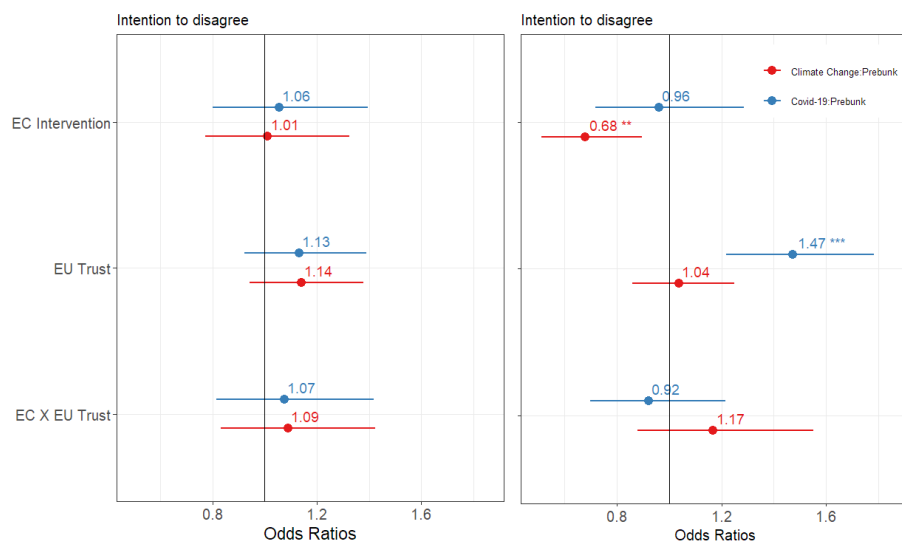


Figure S-14. Treatment effects for average levels of trust in the European Union and interaction between source treatment and trust in the European Union on likelihood to express disagreement with misleading article, by article topic. Shows the estimates for the interaction of the source information and trust in the European Union, both for debunks (left) and prebunks (right), for climate change in red and Covid-19 in blue. EU Trust is demeaned such that a value of 0 on the x-axis corresponds to an average level of trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. The model is a binary logistic regression reporting the odds ratios and heteroscedasticity robust 95%. Significance levels: *** <.001, ** <0.01, * <.05.

Table S-10. Effects of EC-branded vs. neutrally branded interventions, interaction with EU trust and interaction with topic of the misinformation on beliefs and credibility ratings. Shows the estimates for the effects of providing the EC as the source with the neutral-source condition as the baseline. *Models 1 and 2 are ordered logistic regression reporting odds ratios. Models 3 and 4 report linear estimates from OLS models. For all models, heteroscedasticity-robust confidence intervals and p-values are provided.*

Predictors	Agreement with main claim						Credibility assessment					
	Debunk			Prebunk			Debunk			Prebunk		
	Odds Ratios	CI	p	Odds Ratios	CI	p	Estimates	CI	p	Estimates	CI	p
EC (vs. Neutral)	0.88	0.70 – 1.09	0.232	0.97	0.78 – 1.21	0.802	0.01	-0.56 – 0.58	0.982	0.05	-0.53 – 0.64	0.860
EU Trust	0.69	0.57 – 0.85	<0.001	0.55	0.47 – 0.65	<0.001	-0.84	-1.30 – -0.38	<0.001	-1.09	-1.53 – -0.64	<0.001
Climate Change (vs. Covid-19)	0.95	0.76 – 1.18	0.648	1.00	0.80 – 1.24	0.966	0.23	-0.32 – 0.78	0.413	-0.10	-0.67 – 0.46	0.718
EC X EU Trust	0.80	0.62 – 1.04	0.095	0.99	0.78 – 1.25	0.930	-0.51	-1.13 – 0.12	0.111	-0.16	-0.78 – 0.47	0.621
EC X Climate Change	1.15	0.84 – 1.57	0.373	1.11	0.81 – 1.52	0.513	0.04	-0.74 – 0.82	0.921	0.14	-0.67 – 0.95	0.733
EU Trust X Climate Change	1.03	0.79 – 1.33	0.839	1.36	1.08 – 1.71	0.009	0.74	0.12 – 1.35	0.019	0.44	-0.16 – 1.03	0.151
EC X EU Trust X Climate Change	1.07	0.75 – 1.53	0.717	0.83	0.59 – 1.17	0.286	-0.38	-1.24 – 0.48	0.387	0.43	-0.44 – 1.30	0.335
(Intercept)							4.75	4.34 – 5.15	<0.001	4.94	4.51 – 5.36	<0.001
Observations	2066			2012			2066			2012		
R ² Nagelkerke	0.116			0.109			0.042 / 0.039			0.040 / 0.037		

Intercepts for ordered logit model of agreement with the main claim (debunk) are: Strongly disagree | disagree: 0.36, (0.30 – 0.42), p<0.001; Disagree | Neither agree nor disagree: 1.39, (1.18 – 1.63), p<0.001; Neither agree nor disagree | Agree: 4.02, (3.38 – 4.79), p<0.001; Agree | Strongly agree: 15.16, (12.06 – 19.06), p<0.001.

Intercepts for ordered logit model of agreement with the main claim (prebunk) are: Strongly disagree | disagree: 0.36, (0.31 – 0.43), p<0.001; Disagree | Neither agree nor disagree: 1.15, (0.97 – 1.35), p=0.102; Neither agree nor disagree | Agree: 3.12, (2.61 – 3.72), p<0.001; Agree | Strongly agree: 12.62, (10.09 – 15.78), p<0.001.

Table S-11. Effects of EC-branded vs. neutrally branded interventions, interaction with EU trust and interaction with topic of the misinformation on intentions to endorse and criticise. Shows the estimates for the effects of providing the EC as the source with the neutral-source condition as the baseline. All models report odds ratios from binary logistic regressions. For all models, heteroscedasticity-robust confidence intervals and p-values are provided.

Predictors	Intention to agree						Intention to disagree					
	Debunk			Prebunk			Debunk			Prebunk		
	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p
EC (vs. Neutral)	1.20	0.85 – 1.69	0.305	0.98	0.71 – 1.36	0.927	1.06	0.80 – 1.39	0.700	0.96	0.72 – 1.29	0.788
EU Trust	0.67	0.52 – 0.86	0.005	0.71	0.57 – 0.87	0.003	1.13	0.92 – 1.39	0.235	1.47	1.21 – 1.81	<0.001
Climate Change (vs. Covid-19)	0.92	0.65 – 1.31	0.657	0.83	0.60 – 1.16	0.276	1.10	0.83 – 1.44	0.510	1.45	1.10 – 1.92	0.009
EC X EU Trust	1.09	0.78 – 1.53	0.628	0.79	0.58 – 1.07	0.159	1.07	0.81 – 1.42	0.608	0.92	0.69 – 1.23	0.560
EC X Climate Change	0.95	0.58 – 1.55	0.832	1.17	0.73 – 1.86	0.515	0.96	0.65 – 1.41	0.825	0.71	0.47 – 1.05	0.090
EU Trust X Climate Change	1.58	1.11 – 2.25	0.023	1.24	0.90 – 1.70	0.222	1.01	0.76 – 1.33	0.961	0.70	0.53 – 0.93	0.010
EC X EU Trust X Climate Change	0.62	0.39 – 1.01	0.086	1.37	0.87 – 2.16	0.207	1.01	0.68 – 1.50	0.948	1.27	0.85 – 1.91	0.239
(Intercept)	0.17	0.13 – 0.22	<0.001	0.24	0.19 – 0.29	<0.001	0.37	0.30 – 0.45	<0.001	0.33	0.27 – 0.41	<0.001
Observations	2066			2012			2066			2012		
R ² Nagelkerke	0.015			0.024			0.006			0.018		

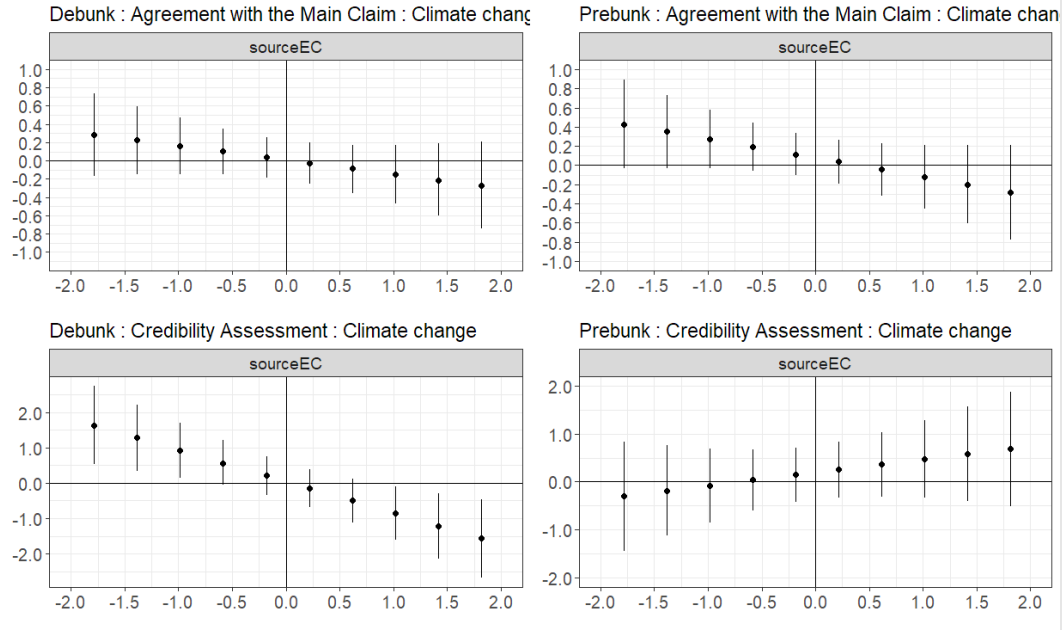


Figure S-15. Effects on beliefs of climate change claims and credibility ratings of EC-branded intervention relative to neutral intervention for different values of EU trust. Middle point is average trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. Shows 95% confidence intervals (not heteroscedasticity robust).

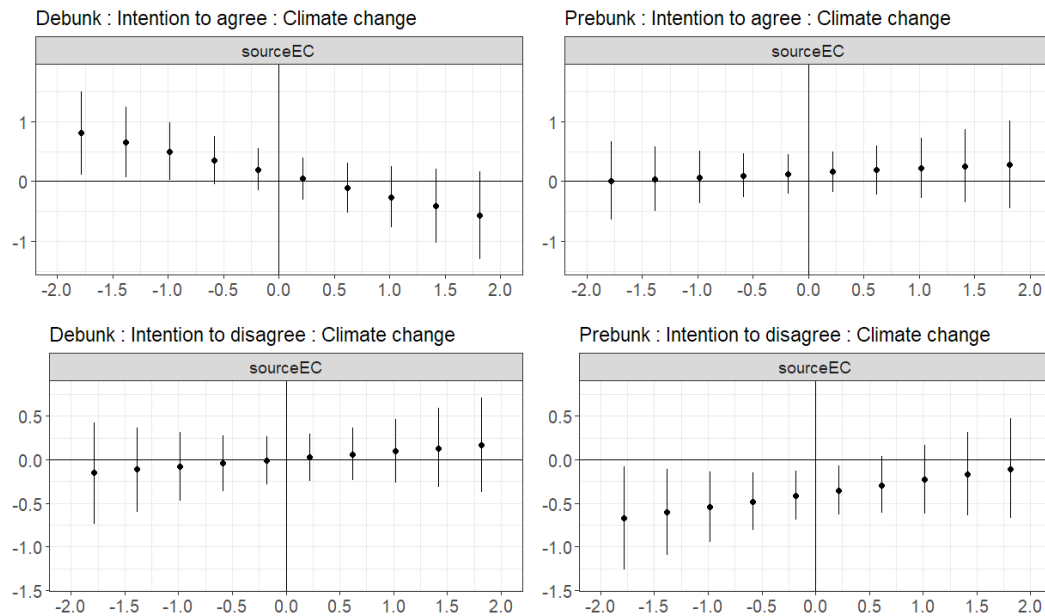


Figure S-16. Effects on intentions to endorse or criticize the misleading climate change article of EC-branded intervention relative to neutral intervention for different values of EU trust. Middle point is average trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. Shows 95% confidence intervals (not heteroscedasticity robust).

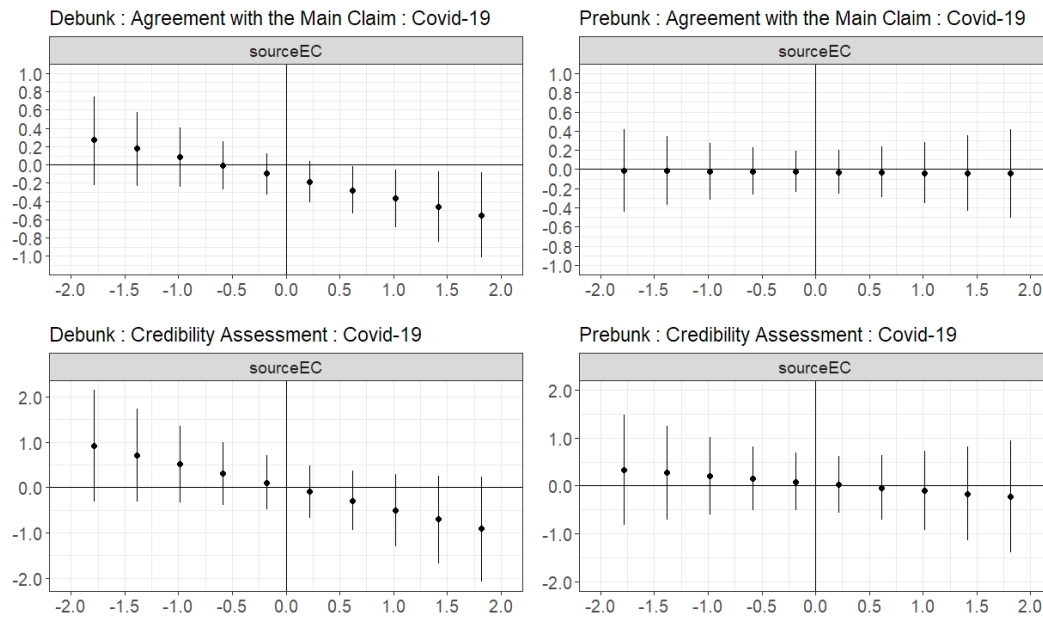


Figure S-17. Effects on beliefs of Covid-19 claims and credibility ratings of EC-branded intervention relative to neutral intervention for different values of EU trust. Middle point is average trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. Shows 95% confidence intervals (not heteroscedasticity robust).

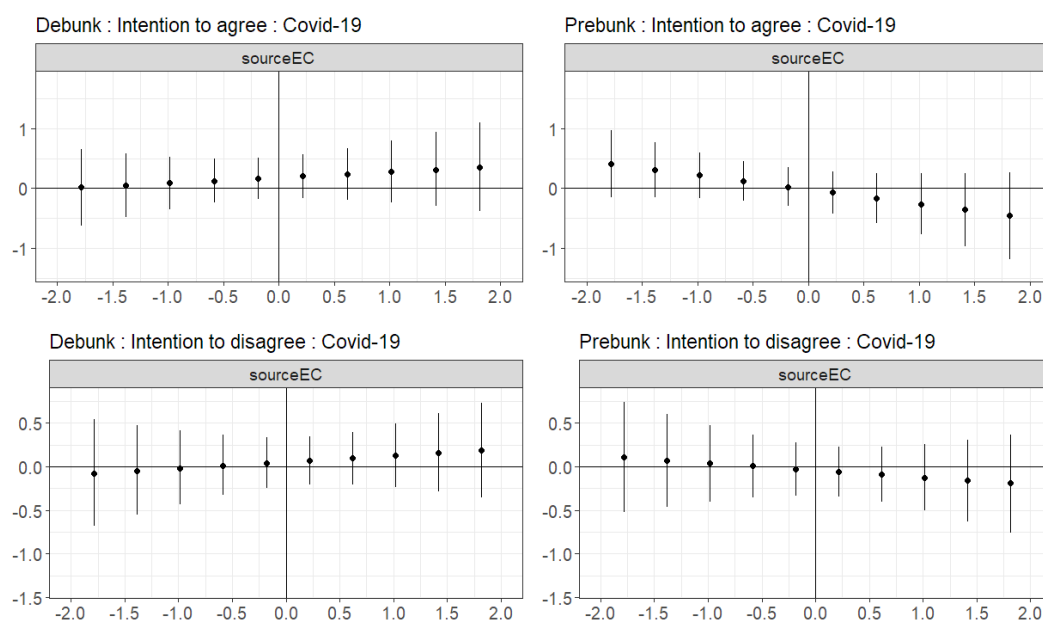


Figure S-18. Effects on intentions to endorse or criticize the misleading Covid-19 article of EC-branded intervention relative to neutral intervention for different values of EU trust. Middle point is average trust in the EU. One unit of change of EU Trust corresponds to one standard deviation. Shows 95% confidence intervals (not heteroscedasticity robust).

Sample characteristics

Table S-12. Sample characteristics by country. Shows the percentages of age, gender and education categories (columns) for the respective countries (rows).

Country	Age						Gender		Education								
	18-24	25-34	35-44	45-54	55-64	>65	Female	Male	Less than primary	Primary	Less than primary	Upper secondary	Upper secondary	Short-cycle tertiary	BA ¹	MA ¹	PhD ¹
Germany	7.63	15.26	15.56	17.01	24.03	20.52	51.26	48.67	0.15	4.88	10.68	35.16	7.4	3.81	19.83	15.64	2.44
Greece	8.99	13.48	16.83	17.97	35.34	7.39	51.94	47.91	0.53	1.75	3.05	21.86	6.47	9.06	41.05	14.01	2.21
Ireland	11.96	16.05	20.68	17.13	18.13	16.05	51.31	46.06	0.15	1	4.32	22.38	13.27	8.8	31.71	16.82	1.54
Poland	8.18	17.05	20.57	15.9	27.91	10.4	52.52	47.32	0.38	2.06	16.36	20.57	9.86	1.83	10.86	37.08	0.99

¹ Or equivalent.

Table S-13. Regional distribution of sample in Germany. Shows percentages of the sample coming from the respective region.

Region	Percentage
Baden-Württemberg	12.28
Bayern	15.71
Berlin	5.11
Brandenburg	3.05
Bremen	0.76
Hamburg	2.44
Hessen	7.78
Mecklemburg	
Vorpommern	1.98
Niedersachsen	9.99
Nordrhein Westfalen	21.21
Rheinland Pfalz	4.58
Saarland	1.22
Sachsen	5.11
Sachsen Anhalt	2.82
Schleswig Holstein	3.36
Thüringen	2.59

Table S-14. Regional distribution of sample in Greece. Shows percentages of the sample coming from the respective region.

Region	Percentage
Attica	41.81
Central Greece	2.67
Central Macedonia	20.56
Crete	4.04
Eastern Macedonia and Thrace	5.41
Epirus	2.97
Ionian Islands	1.29
North Aegean	1.29
Peleponnese	6.78
South Aegean	2.21
Thessaly	5.71
Western Greece	2.67
Western Macedonia	2.59

Table S-15. Regional distribution of sample in Ireland. Shows percentages of the sample coming from the respective region.

Region	Percentage
Eastern and Midland	42.44
Northern and Western	20.45
Southern	37.11

Table S-16. Regional distribution of sample in Poland. Shows percentages of the sample coming from the respective region.

Region	Percentage
Dolnośląskie	6.27
Kujawsko-pomorskie	5.89
Łódzkie	6.5
Lubelskie	6.12
Lubuskie	1.91
Małopolskie	7.87
Mazowieckie: pozostałe Powiaty	6.73
Mazowieckie: Warszawa, warszawski wschodni, warszawski zachodni	8.87
Opolskie	2.14
Podkarpackie	4.82
Podlaskie	3.52
Pomorskie	5.73
Śląskie	13.69
Świętokrzyskie	2.91
Warmińsko-mazurskie	2.91
Wielkopolskie	9.4
Zachodniopomorskie	4.74

Robustness checks

Table S-17. Results from models for the four main outcome variables including control variables. Shows the estimates for the intervention effects with the Control condition as the baseline. *Model 1 is an ordered logistic regression reporting the odds ratios. Model 2 reports linear estimates from an OLS model. Models 3 and 4 report odds ratios from a binary logistic regression. For all models, heteroscedasticity-robust confidence intervals and p-values are provided.*

Predictors	Agreement with the Main Claim			Credibility Assessment			Intention to agree			Intention to disagree		
	Odds Ratios	CI	p	Estimates	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p
Neutral Debunk	0.44	0.36 – 0.53	<0.001	-1.53	-1.99 – -1.07	<0.001	0.43	0.32 – 0.59	<0.001	1.39	1.08 – 1.79	0.010
EC Debunk	0.40	0.33 – 0.49	<0.001	-1.52	-1.96 – -1.08	<0.001	0.53	0.40 – 0.70	<0.001	1.44	1.14 – 1.82	0.003
Neutral Prebunk	0.56	0.45 – 0.68	<0.001	-1.13	-1.60 – -0.67	<0.001	0.60	0.44 – 0.80	0.001	1.39	1.08 – 1.79	0.010
EC Prebunk	0.54	0.45 – 0.65	<0.001	-1.28	-1.70 – -0.85	<0.001	0.67	0.51 – 0.87	0.003	1.04	0.83 – 1.32	0.728
EU Trust	0.99	0.89 – 1.09	0.794	0.07	-0.17 – 0.30	0.581	1.11	0.96 – 1.27	0.165	1.10	0.98 – 1.24	0.130
Intercept				10.46	8.92 – 12.00	<0.001	1.09	0.42 – 2.81	0.861	0.15	0.07 – 0.35	<0.001
Control variables		Yes			Yes			Yes			Yes	
Observations		4562			4562			4562			4562	
R ² Nagelkerke		0.533			0.233 / 0.226			0.095			0.035	

Intercepts for ordered logit model of agreement with the main claim are: Strongly disagree | disagree: 0.02, (0.01 – 0.04), p<0.001; Disagree | Neither agree nor disagree: 0.09, (0.05 – 0.18), p<0.001; Neither agree nor disagree | Agree: 0.28, (0.14 – 0.55), p<0.001; Agree | Strongly agree: 1.15, (0.59 – 2.26), p=0.675.

Table S-18. Effects of EC-branded vs. neutrally branded interventions and interaction with EU trust on beliefs and credibility ratings including control variables. Shows the estimates for the effects of providing the EC as the source with the neutral-source condition as the baseline. *Models 1 and 2 are ordered logistic regression reporting the odds ratios. Models 3 and 4 report linear estimates from OLS models. For all models, heteroscedasticity-robust confidence intervals and p-values are provided*

Intervention	Agreement with the Main Claim						Credibility Assessment					
	Debunk			Prebunk			Debunk			Prebunk		
Predictors	Odds Ratios	CI	p	Odds Ratios	CI	p	Estimates	CI	p	Estimates	CI	p
Intercept							11.27	8.81 – 13.73	<0.001	7.85	5.43 – 10.27	<0.001
EC (vs. Neutral)	0.93	0.78 – 1.11	0.439	1.00	0.83 – 1.21	0.985	0.06	-0.32 – 0.45	0.752	-0.18	-0.58 – 0.23	0.386
EU Trust	1.05	0.86 – 1.28	0.620	1.12	0.94 – 1.34	0.204	0.31	-0.11 – 0.72	0.148	0.06	-0.34 – 0.47	0.752
EC Intervention X EU Trust	0.87	0.71 – 1.07	0.192	0.88	0.73 – 1.07	0.200	-0.51	-0.94 – -0.08	0.019	0.18	-0.24 – 0.60	0.398
Control variables	Yes			Yes			Yes			Yes		
Observations	1818			1801			1818			1801		
R2 Nagelkerke	0.540			0.517			0.235 / 0.217			0.262 / 0.245		

Intercepts for ordered logit model of agreement with the main claim (debunk) are: Strongly disagree | disagree: 0.02, (0.01 – 0.06), p<0.001; Disagree | Neither agree nor disagree: 0.10, (0.03 – 0.31), p<0.001; Neither agree nor disagree | Agree: 0.32, (0.10 – 1.02), p=0.054; Agree | Strongly agree: 1.27, (0.40 – 4.04), p=0.688.

Intercepts for ordered logit model of agreement with the main claim (prebunk) are: Strongly disagree | disagree: 0.04, (0.01 – 0.12), p<0.001; Disagree | Neither agree nor disagree: 0.16, (0.05 – 0.46), p=0.001; Neither agree nor disagree | Agree: 0.48, (0.17 – 1.40), p=0.178; Agree | Strongly agree: 2.07, (0.71 – 6.05), p=0.181.

Table S-19. Effects of EC-branded vs. neutrally branded interventions and interaction with EU trust on intentions to agree and disagree. Shows the estimates for the effects of providing the EC as the source with the neutral-source condition as the baseline. All models report odds ratios from binary logistic regressions. For all models, heteroscedasticity-robust confidence intervals and p-values are provided.

Intervention	Intention to agree						Intention to disagree					
	Debunk			Prebunk			Debunk			Prebunk		
Predictors	Odds Ratios	CI	p	Odds Ratios	CI	p	Estimates	CI	p	Estimates	CI	p
Intercept	1.00	0.19 – 5.27	0.997	0.33	0.07 – 1.55	0.163	0.32	0.09 – 1.17	0.083	0.13	0.03 – 0.49	0.002
EC (vs. Neutral)	1.23	0.93 – 1.63	0.164	1.07	0.82 – 1.42	0.619	1.04	0.84 – 1.29	0.713	0.70	0.56 – 0.89	0.003
EUTrust	1.14	0.86 – 1.50	0.388	1.16	0.90 – 1.50	0.261	1.06	0.86 – 1.32	0.586	1.08	0.87 – 1.34	0.498
EC X EU Trust	0.89	0.68 – 1.16	0.427	0.92	0.71 – 1.18	0.544	1.11	0.89 – 1.37	0.355	1.08	0.86 – 1.34	0.511
Control variables	Yes			Yes			Yes			Yes		
Observations	2066			2012			2066			2012		
R2 Nagelkerke	0.011			0.016			0.006			0.012		

Experimental materials

The Covid-19 vaccine is dangerous

It's obvious and the scientific evidence is clear: the Covid-19 vaccine is dangerous, scientific study proves.

Written by [REDACTED]
February 10, 2022



A scientific paper that was recently published in a leading academic journal proves that the Covid-19 vaccine is dangerous for people's health.

The article proves that the vaccine's side effects, such as an elevated risk for thrombosis and heart muscle inflammations, are more harmful to the body than the virus itself.

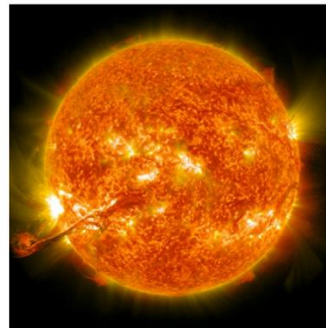
This exposes previous "scientific" evidence claiming the contrary. It also proves that so-called experts lied, probably because they were paid by Big Pharma.

The claim that the Covid-19 vaccine is safe is not just illogical, but also an immoral lie to the face of the people. The evidence clearly proves what many of us have been guessing for a long time: the Covid-19 vaccine is dangerous.

The sun, not humans, causes climate change

It's obvious and the scientific evidence is clear: the sun, not humans, causes climate change, scientific study proves.

Written by [REDACTED]
February 10, 2022



A scientific paper that was recently published in a leading academic journal proves that the sun, not humans, causes climate change.

The article proves that recent warming is caused through changes in solar irradiance, meaning the amount of solar energy that reaches the Earth. This is a completely natural process.

This exposes previous "scientific" evidence claiming the contrary. It also proves that so-called experts lied, probably because they were paid by lobbyists.

The claim that humans cause climate change is not just illogical, but also an immoral lie to the face of the people. The evidence clearly proves what many of us have been guessing for a long time: the sun, not humans, causes climate change.

Figure S-19. Examples of two misleading articles used in the experiment. Left: Covid-19. Right: Climate change.

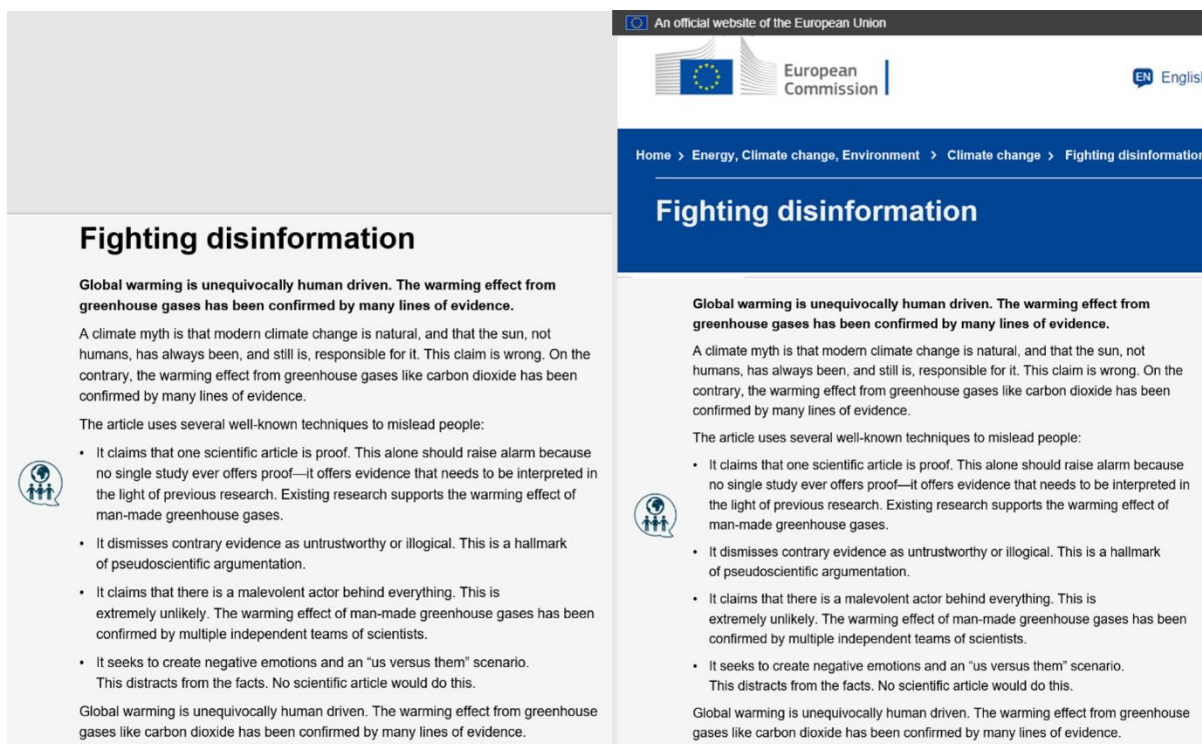


Figure S-20. Examples of two debunks used in the experiment. Left: no source. Right: EC-source.

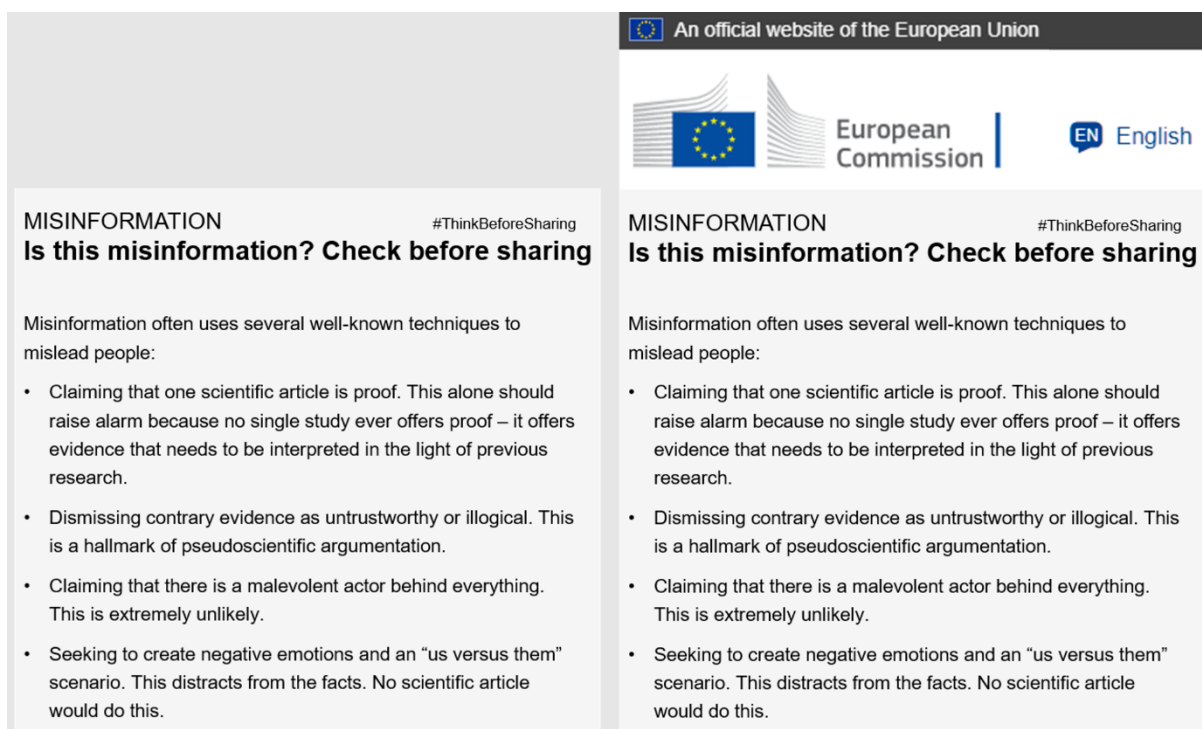


Figure S-21. Examples of prebunks used in the experiment. Left: no source. Right: EC-source.

Questionnaire

1. We will now ask you about your thoughts and feelings regarding the European Union. In this context, the “European Union” refers to its main institutions, which are the European Commission, the European Parliament, and the European Council. [5 Point Likert, 1-Strongly disagree – 5-Strongly agree, Prefer not to say]
 1. Overall, the EU is competent and efficient
 2. The EU usually carries out its duties poorly (reverse coded)
 3. The EU usually acts in its own interests (reverse coded)
 4. The EU wants to do its best to serve Europe
 5. The EU is generally free of corruption
 6. The EU work is open and transparent
2. Please respond to the following questions: [10 Point, 1-I do not trust it at all – 10-I trust it completely, Prefer not to say]
 - How much trust do you have in the European Union?
 - How much trust do you have in your national government?
3. Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people? [3 Options: Most people can be trusted, Need to be very careful, Prefer not to say]
4. Please indicate to what extent you agree or disagree with the following statement: [10 Point, 1-Left – 10-Right, Prefer not to say]
 - In political matters people talk of 'the left' and 'the right'. How would you place your views on this scale?
5. Please indicate to what degree you agree or disagree with the following statements concerning the CONTROL/PREBUNK/DEBUNK text that you have seen before (shown again below). [5 Point Likert, 1-Strongly disagree – 5- Strongly agree, Prefer not to say]
 1. The message appears relevant to me
 2. I can use this message to make good decisions
 3. The message appears authentic to me
 4. The message grabbed my attention
 5. The message wants to manipulate me
6. Who do you think was its source? [4 Options: No one, The European Commission, The University of Hamburg, I don't know]
7. How often, on average, do you use online social media (e.g., Facebook, Twitter, Instagram, TikTok, etc.) [5 Options: Seldom or never, Several times a month, At least once a week, Every day or almost every day, Prefer not to say]
8. How often do you come across news or information that you believe misrepresent reality or are even false? [5 Options: Seldom or never, Several times a month, At least once a week, Every day or almost every day, Prefer not to say]
9. How important is it to you that you only share news articles on social media e.g., Facebook, Twitter, Instagram, TikTok, etc.) if they are accurate? [5 Point Likert, 1-Very unimportant – 5-]
10. How confident are you that you are able to identify news or information that misrepresent reality or are even false? [4 Point Likert, 1-Not at all confident – Very confident, Prefer not to say]
11. Please indicate the degree to which the following statements are characteristic of you: [5 Point Likert, 1-Extremely uncharacteristic of me – 5-Extremely characteristics of me, Prefer not to say]

1. I like to have the responsibility of handling a situation that requires a lot of thinking
2. I would prefer complex to simple problems
3. Thinking is not my idea of fun
4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities
5. I really enjoy a task that involves coming up with new solutions to problems
6. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought

Debriefing text

For people in the Control and Prebunk condition:

Please be aware that the first article titled [ARTICLE TITLE] you saw previously was fabricated and contained incorrect information. Please carefully read the following correction. After reading, check the box indicating that you read the article before advancing.

[SHOW CORRESPONDING DEBUNK]

For everyone:

Thank you for taking the time to respond to this survey. The goal of this study was to find out how effective different ways to expose and correct (debunk) false information (misinformation, fake news) are. To investigate this, we asked you to read four articles which contained false claims and could thus be considered misinformation or fake news. At some point, you were then shown four articles correcting these false claims and explaining the deceptive strategies used in them. The nature of the phenomenon we are investigating required minor deception on our part. Specifically, we presented the fake news articles without labelling them as such. In this way, we may have led you to believe them to be accurate. To investigate the effectiveness of correcting (debunking) information, there was no other way than to expose you to fake news and only correct them at a later point in time. This is sometimes necessary in this type of research. If we tell people about the articles being fake in advance, we could not investigate how debunks work for people who encounter fake news without realizing it. Your participation is greatly appreciated by the researchers involved and will contribute to advancing the research in this field. If you have any questions about this study, please contact us. Finally, we urge you not to discuss this study with anyone else who is currently participating or might participate at a future point in time. As you can certainly appreciate, we will not be able to examine the effectiveness of correcting and debunking misinformation for participants who know about the true purpose of the project beforehand. Thank you!