nature human behaviour

Article

https://doi.org/10.1038/s41562-024-02023-2

Inoculation and accuracy prompting increase accuracy discernment in combination but not alone

Received: 21 August 2023

Accepted: 20 September 2024

Published online: 04 November 2024

Check for updates

Gordon Pennycook[1,2] ✉, Adam J. Berinsky[3], Puneet Bhargava[4,5], Hause Lin[2,6], Rocky Cole[7], Beth Goldberg[7], Stephan Lewandowsky[8,9,10] & David G. Rand[6,11,12]

Misinformation is a major focus of intervention efforts. Psychological inoculation—an intervention intended to help people identify manipulation techniques—is being adopted at scale around the globe. Yet the efficacy of this approach for increasing belief accuracy remains unclear, as prior work uses synthetic materials that do not contain claims of truth. To address this issue, we conducted five studies with 7,286 online participants using a set of news headlines based on real-world true/false content in which we systematically varied the presence or absence of emotional manipulation. Although an emotional manipulation inoculation did help participants identify emotional manipulation, there was no improvement in participants' ability to tell truth from falsehood. However, when the inoculation was paired with an intervention that draws people's attention to accuracy, the combined intervention did successfully improve truth discernment (by increasing belief in true content). These results provide evidence for synergy between popular misinformation interventions.

Global concerns about the spread of misinformation and its toxic effects on democracy have led to a surge of interest in potential mitigation measures[1–6]. One of the most common approaches is to undermine the influence of falsehoods by providing countervailing information. For example, to ensure that people do not believe misinformation (defined here as information that is false or misleading, whatever the intent), research has focused on assessing the efficacy of fact checks, corrections and 'debunks'[7–10]. Although these approaches show promise—corrections do tend to undermine the effect of misinformation and rarely backfire[11]—they are loss-mitigation strategies at best. Corrections necessarily occur after misinformation has already spread online, and thus they should be only one element of a broader strategy.

Researchers and practitioners have therefore also focused on strategies that attempt to get ahead of misinformation and limit its potential influence before people are exposed. Recent research demonstrates that 'prebunking' misinformation by providing corrective information in anticipation of falsehoods can protect people against being misled[12–14]. One promising prebunking strategy that has been gaining popularity is attitudinal or psychological 'inoculation'. The goal of inoculation is to provide individuals with knowledge or abilities that help them spot misinformation by exposing them to a weakened

[1]Department of Psychology, Cornell University, Ithaca, NY, USA. [2]Hill/Levene Schools of Business, University of Regina, Regina, Saskatchewan, Canada. [3]Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA. [5]Department of Marketing, Wharton School, University of Pennsylvania, Philadelphia, PA, USA. [6]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Jigsaw (Google), Mountain View, CA, USA. [8]School of Psychological Science, University of Bristol, Bristol, UK. [9]Department of Psychology, University of Potsdam, Potsdam, Germany. [10]School of Psychological Science, University of Western Australia, Crawley, Western Australia, Australia. [11]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [12]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: gordon.pennycook@cornell.edu

form of a misleading technique (or set of techniques)[14–16]. Critically, unlike other prebunking techniques, inoculation techniques are not intended to be specific to unique falsehoods but rather are intended to boost people's general ability to identify misleading techniques that are thought to be common to misinformation[17,18]. Note that psychological inoculation—and inoculation theory in general—have been applied very broadly in the persuasion literature[17,18]; however, we focus here on the use of the intervention as a tool to lower people's susceptibility to misinformation in particular.

To illustrate, a recent set of experiments have tested a scalable version of psychological inoculation. Roozenbeek et al.[15] developed a set of brief videos (less than two minutes each) that teach users about manipulation techniques that are thought to be statistical markers of misinformation (namely, emotional language, ad hominem attacks, false dichotomies, incoherence and scapegoating). For example, the emotional manipulation inoculation video informs users about how emotional language (such as words like 'disgusting' and 'horrifying') can be used to manipulate users into paying attention to content online (all videos are accessible at https://inoculation.science/). These videos, generally speaking, improved participants' ability to discern whether there was manipulation in the synthetic tweets that the researchers created to either contain the technique or not—with the emotional language inoculation having one of the clearest effects on the ability to discern between manipulative and neutral content.

Two of the inoculation videos were also tested in a quasi-field experiment on YouTube (including the emotional language inoculation that we focus on here). For this, users were presented an inoculation video in the standard course of their YouTube usage as an ad and then were asked to identify the manipulation technique in a multiple-choice question. The researchers found that those who viewed the inoculative video were 5–6% more likely to correctly identify a manipulation technique relative to a control group, suggesting that a short video on social media may be sufficient to teach people how to identify misinformation techniques[15]. Efforts by technology companies to scale this approach are already underway[19–21].

Although inoculation thus shows considerable promise as a scalable approach, research on inoculation has focused on demonstrating people's ability to detect when they might be misled by boosting their ability to recognize misleading or deceptive argumentation[15,22]. This is a valuable skill, in particular because it provides protection against poor argumentation irrespective of the specific content. For example, the use of highly emotive or inflammatory language may signal manipulative intent irrespective of whether the message is entirely false, contains seeds of truth or is accurate.

However, it remains unclear whether inoculation also makes people better at discerning the accuracy of claims—that is, whether inoculation increases the ability to differentiate true from false information. Past work has primarily tested the efficacy of inoculation by using content that contains the targeted technique (such as emotional manipulation) but that does not necessarily contain specific truth claims (or claims that are evaluable as true or false). For example, Roozenbeek et al.[15] used the following as an example of emotional language manipulation: "What this airline did for its passengers will make you tear up - So heartwarming" (with no additional context or information about the airline or what they actually did). Items such as this are intended to be 'pure' examples of the technique; naturally, however, the missing context will be present when such interventions are used in the real world, and it is important to ascertain whether technique identification skills extend to recognizing the truth or falsehood of statements.

Typically, participants are also explicitly asked about the presence/absence of the technique instead of the accuracy of the statements (but see ref. 23). It is thus unclear whether boosting technique recognition carries over to affect judgements about the truth or falsity of true and false content. Does teaching about misleading techniques (on its own) help people detect misinformation? And if not, what can be done to address this issue?

Notwithstanding the promise of inoculation, there are also reasons to believe that inoculations with a specific focus on misleading techniques (such as using emotional language) may not necessarily carry over to improve judgements about the accuracy of true or false content. In particular, research has shown that people are often inattentive to accuracy on social media[24–26]. For example, in many experiments when participants are asked directly to assess whether they believe a set of true and false news headlines, most people are actually quite good at distinguishing between them (that is, they believe the true news much more than the false news); however, if participants are instead asked to indicate whether they would share the news on social media, they are largely insensitive to the veracity of the headlines[24–26]. In fact, the social media context may itself be responsible for distracting people from thinking about accuracy, as simply having participants make sharing judgements reduces their ability to tell true from false when making accuracy judgements[27].

One approach to overcoming this inattentiveness to accuracy is through very brief interventions that prompt the reader to consider accuracy, known as 'accuracy prompts' (or 'accuracy primes' or 'accuracy nudges'). These prompts have demonstrated promise in improving people's attentiveness to accuracy when making judgements about what to share online (that is, by increasing the quality of content that they share)[25,26,28] (for a meta-analysis, see ref. 29). This is thought to occur because the prompts redirect participants' attention to the concept of accuracy, and most participants do not wish to share content that they realize is inaccurate[24,26]. Thus, simply considering accuracy reduces the (undesired) sharing of content that participants would have been able to identify as inaccurate had they thought about it[30].

It is therefore possible that the effects of inoculations that focus only on identifying manipulation techniques—but do not address the problem of trying to specifically spot false claims—may not transfer to improved misinformation detection. However, if before delivering the inoculation, the intervention explicitly draws people's attention to accuracy and the need to differentiate truth from falsehood, this juxtaposition may induce transfer and successfully lead to improved truth discernment. Put differently, there may be a synergy between inoculation approaches (which teach people specific information about manipulation techniques but may not focus their attention on detecting falsehoods) and accuracy prompt approaches (which direct attention to truth and falsity but do not actually teach individuals anything about how to detect falsehoods). Thus, although accuracy prompts are not intended to improve people's ability to distinguish between true and false content—as participants are simply having their attention directed to accuracy without being taught anything new—inducing an accuracy mindset could improve the effectiveness of psychological inoculations.

Here we investigate whether psychological inoculation effects carry over to improved truth discernment, as well as the potential synergy between inoculation and accuracy prompting. We show in three experiments that, while inoculation reliably boosts people's ability to detect problematic content, it does not by itself increase truth discernment. However, we show in two further experiments that when inoculation is combined with a prompt that reminds people of the problem of misinformation and the importance of accuracy, the combined intervention (but neither prompt nor inoculation in isolation) boosts participants' accuracy discernment.

To assess participants' ability to identify inaccurate claims that systematically vary in the presence of a manipulation technique, we developed a set of stimuli that were sourced from real-world true and false news headlines but that were modified to contain either the manipulation technique of emotional language or no such technique. For this, we took a larger corpus of 83 political headlines (43 false, 40 true) and created two versions for each claim: one intended to elicit

Table 1 | Average emotionality rating in the pretest for matched high- versus low-emotionality headlines

	Target	Low emotionality	High emotionality	Difference	*t*-test	Cohen's *d*
False	Democratic	3.0	3.7	0.7	t_7=19.9, P<0.001	7.02
	Republican	2.8	3.3	0.5	t_7=10.4, P<0.001	3.68
True	Democratic	3.0	3.6	0.6	t_7=6.7, P<0.001	2.37
	Republican	3.1	3.6	0.5	t_7=4.4, P=0.003	1.54

Displayed are in-party ratings—that is, emotionality ratings among Democrats for Democratic-targeted headlines and among Republicans for Republican-targeted headlines. The tests are two-sided.

strong emotions from the reader by including evocative language and one that did not include an emotional manipulation (and that was also stripped of evocative language). For example, the following is a false high-emotion headline: 'Decorated Democrat arrested and charged with 6 heinous crimes'. The non-emotion version of the same headline was: 'A Democrat has been arrested and charged with 6 crimes'. We then took this corpus and completed a pretest where participants (n = 982) rated the extent to which a random subset of 15 headlines made them feel emotions (for example, angry, sad, worried, happy or excited) using a six-point scale from 'not at all' to 'extremely'. The participants also rated the headlines in terms of their perceived likelihood and (Democratic versus Republican) partisanship, and indicated whether they would share the headline on social media. From this, we selected 32 headlines where the high-emotion version elicited particularly strong emotional reactions relative to the low-emotion version (Table 1). This procedure gave us a set of headlines that not only differed in terms of whether they contained the emotional manipulation technique (as in past work) but also were pretested to be significantly different in terms of the actual emotions that they elicited. Although adding emotional language to the headlines may in theory subtly shift the claims (see https://osf.io/ym5wg/ for the full materials), here we were able to do so without changing the veracity of the headlines (that is, while adding emotional language to true headlines may subtly influence their plausibility, they did not become false—and removing emotional language from false headlines did not make them true). We also included a question about the perceived plausibility of the headlines in the pretest, and the high- versus low-emotionality headlines did not significantly differ at the item level (Supplementary Section 5). We did not include information about the publisher of the headlines, as they were changed to either add or remove emotional language, and thus none of the headlines were, strictly speaking, used by any publishers. In any case, past work has shown that hiding versus revealing publisher information on headlines has little effect on judgements of accuracy[31].

This pretested set of news headlines allows us to test whether the emotional manipulation inoculation is successful in helping people identify misinformation using real-world stimuli that contain true or false claims. In addition, in the studies below, the participants were asked about manipulativeness and belief (accuracy) separately so that we could investigate whether participants spontaneously identified misinformation to a greater extent after the intervention. Importantly, following recent recommendations[32], we focused on whether the intervention improves both truth and technique discernment (that is, whether participants judge true headlines as more accurate than false headlines, and whether they judge high-emotion headlines as less accurate than low-emotion headlines). Researchers have raised concerns that psychological inoculation may increase general scepticism[33], and, ultimately, if the inoculation is improving people's ability to detect emotional language manipulation, it should result in an improved ability to distinguish between items that do versus do not contain manipulation (for further discussion, see ref. 32).

Because we orthogonally manipulated both the facticity and the emotionality of the stimuli, we were able to assess two versions of discernment. First, do users distinguish between true and false content; that is, are ratings of belief (manipulativeness) lower (higher) for false than for true news headlines? Second, do users distinguish between emotionally manipulative (high emotion) and not emotionally manipulative (low emotion) content; that is, are ratings of belief (manipulativeness) lower (higher) for manipulative than for non-manipulative news headlines?

Results

In Experiment 1, the participants were randomized into an emotional inoculation treatment condition or one of three control conditions. As noted, we used a set of real-world true and false headlines (16 of each) and modified them by including emotionally charged language (primarily fear or anger evoking, given the content). The participants saw only one version (high versus low emotionality) of each headline (16 in total), and this presentation was counterbalanced across participants. After viewing the randomly selected headlines, the participants rated the trustworthiness and the manipulativeness of each headline on a six-point scale: (1) extremely untrustworthy to (6) extremely trustworthy, and (1) not at all manipulative to (6) extremely manipulative. The full materials, data and preregistrations can be found on https://osf.io/ym5wg/; the inoculation video is available at https://inoculation.science/. See Methods for the full description. All statistical tests are two-tailed.

We ran two linear regression models with robust standard errors clustered on subject and headline, predicting (a) manipulativeness and (b) trustworthiness using an inoculation treatment dummy (0 = control, 1 = treatment), a news headline veracity dummy (−0.5 = false, 0.5 = true), an emotional manipulativeness dummy (−0.5 = not emotional, 0.5 = emotional) and their interactions (Supplementary Table 1). For these models, we report regression coefficients, b, with 95% confidence intervals. First, supporting the validity of our items, true content was judged as less manipulative (b = −1.24 (−1.60, −0.88), z = −6.79, P < 0.001) and more trustworthy (b = 1.46 (1.16, 1.76), z = 9.66, P < 0.001) than false content in the control condition, and high-emotion content was judged as more manipulative (b = 0.46 (0.35, 0.57), z = 8.10, P < 0.001) and less trustworthy (b = −0.28 (−0.38, −0.18), z = −5.35, P < 0.001) than low-emotion content in the control condition—although judgements of manipulativeness/trustworthiness were apparently much more influenced by truth than by emotionality. Second, replicating Roozenbeek et al.[15], we found that the inoculation video significantly increased the difference in manipulativeness and trustworthiness judgements for low- versus high-emotion content—that is, the treatment increased emotion discernment (manipulativeness: b = 0.42 (0.31, 0.53), z = 7.75, P < 0.001; trustworthiness: b = −0.22 (−0.27, −0.17), z = −8.50, P < 0.001; Fig. 1). In particular, the inoculation video significantly increased the perceived manipulativeness of high-emotion headlines (b = 0.18 (0.07, 0.29), z = 3.09, P = 0.002, d = 0.10) but did not significantly decrease the trustworthiness of high-emotion headlines (b = −0.06 (−0.13, 0.02), z = −1.50, P = 0.134, d = −0.04). The inoculation also significantly decreased the perceived manipulativeness of low-emotion headlines (b = −0.24 (−0.35, −0.14), z = −4.48, P < 0.001, d = −0.14) and increased the trustworthiness of low-emotion headlines (b = 0.16 (0.09, 0.24), z = 4.19, P < 0.001, d = 0.10).

We next asked whether there was a carry-over effect of the inoculation to news headline truth discernment. That is, did the inoculation

https://doi.org/10.1038/s41562-024-02023-2

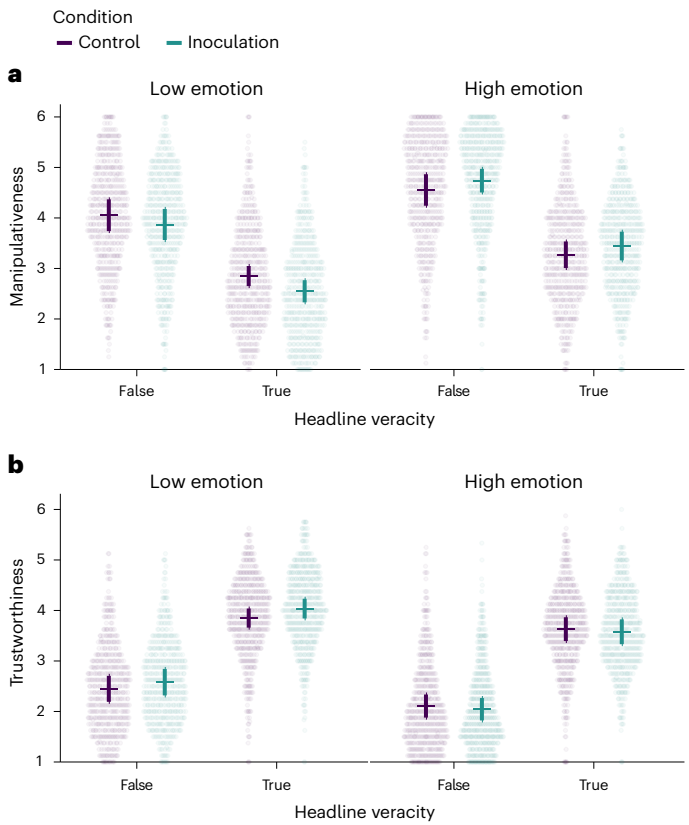

Fig. 1 | Experiment 1 ratings by condition, headline veracity and emotional manipulativeness. a,**b**, Average adjusted predictions (horizontal lines) and 95% confidence intervals (vertical lines) are shown for manipulativeness of headlines (**a**) and trustworthiness of headlines (**b**). Each dot is one participant's mean rating (n = 1,030). The predictions were obtained from ordinary least squares (OLS) regression analysis with robust standard errors clustered on participant and item (two-sided tests without adjustments for multiple comparisons).

treatment interact significantly with veracity (either directly or in a three-way interaction with emotional manipulativeness)? The answer in both cases is no (Supplementary Table 2). The inoculation did not significantly interact with news headline veracity for either manipulativeness (b= −0.07 (−0.20, 0.07); z = −0.97; P = 0.332; Bayes factor (BF), <0.01) or trustworthiness (b = 0.03 (−0.06, 0.13), z = 0.69, P = 0.488, BF < 0.01), nor was there a three-way interaction between the inoculation treatment, headline veracity and emotionality for manipulativeness (b = 0.11 (−0.10, 0.31), z = 1.03, P = 0.303, BF < 0.01) or trustworthiness (b = −0.04 (−0.14, 0.05), z = −0.89, P = 0.372, BF < 0.01). Thus, although the inoculation facilitated the detection of emotional manipulativeness in news headlines, this effect was equivalent for true and false headlines. Furthermore, we did not find evidence that teaching people about emotional manipulativeness (by itself) affected whether they could distinguish between true and false news headlines. Also, as noted, even judgements of manipulativeness were influenced more than twice as strongly by truth as by emotionality.

One issue, however, is that the identification of the manipulativeness technique in Experiment 1 was not 'spontaneous' identification in the sense that the participants were given the explicit task of searching for manipulativeness and trustworthiness. This may have distracted them from focusing on spotting misinformation and therefore may have blunted any carry-over effect of the inoculation on truth discernment. Thus, in Experiment 2, we used the same design as in Experiment 1 but instead asked the participants to judge the accuracy of the claims contained in the news headlines (and therefore not manipulativeness per se). This contributes valuable understanding of how inoculating against a manipulative technique carries over to accuracy judgements or not.

We ran a linear regression model with robust standard errors clustered on subject and headline, predicting accuracy judgements using an inoculation treatment dummy (0 = control, 1 = treatment), a news headline veracity dummy (−0.5 = false, 0.5 = true), an emotional manipulativeness dummy (−0.5 = not emotional, 0.5 = emotional) and their interactions (Supplementary Table 2). Similar to Experiment 1, true content was judged as more accurate than false content in the control condition (b = 0.38 (0.32, 0.44), z = 11.93, P < 0.001), and high-emotion content was judged as less accurate than low-emotion content in the control condition (b = −0.05 (−0.07, −0.03), z = −4.83, P < 0.001). Furthermore, the inoculation video significantly increased the difference in perceived accuracy for low- versus high-emotion content—that is, the treatment increased emotion discernment (b = −0.02 (−0.03, −0.01), z = −3.20, P = 0.001). However, as in Experiment 1, this effect was fairly small (Fig. 2). In particular, the inoculation video significantly decreased the perceived accuracy of high-emotion headlines (b = −0.01 (−0.03, 0.00), z = −1.97, P = 0.049, d = −0.04) but had no significant effect on the perceived accuracy of low-emotion headlines (b = 0.01 (−0.01, 0.02), z = 0.94, P = 0.345, d = 0.02).

The inoculation treatment did not, however, have a significant carry-over effect on truth discernment (Supplementary Table 2). Specifically, there was no interaction between the inoculation and news headline veracity (b = −0.01 (−0.03, 0.01), z = −0.76, P = 0.446, BF = 0.007), nor was there a three-way interaction between the inoculation treatment, headline veracity and emotionality (b = −0.001 (−0.02, 0.02), z = −0.10, P = 0.919, BF = 0.006).

As a follow-up, we also restricted the analysis to only false high-emotion and true low-emotion items; this represents a maximally favourable limiting case where emotional manipulation is always present in false content and never present in true content. Thus, in this case, decreasing belief in emotional content should be expected to also decrease belief in false content (and therefore increase discernment). However, even here, we did not find a significant effect of the inoculation on truth discernment (interaction between condition and veracity, b = 0.01 (−0.01, 0.04), z = 0.83, P = 0.406, BF = 0.011). Accordingly, there was no significant inoculation effect on accuracy judgements for either false high-emotion (b = −0.01 (−0.03, 0.01), z = −0.84, P = 0.401, d = −0.03, BF = 0.016) or true low-emotion (b = 0.01 (−0.01, 0.02), z = 0.23, P = 0.815, d = 0.01, BF = 0.011) headlines.

Experiment 2 confirmed that the inoculation video decreased belief in emotionally manipulative content relative to non-emotionally manipulative content. Thus, the results of Experiment 1 (and those of Roozenbeek et al.[15]) do not appear to be driven entirely by explicitly focusing participants on the task of identifying manipulativeness (by asking about it directly instead of asking about accuracy, as we did here—of course, the participants were nonetheless asked directly about accuracy; see Supplementary Section 3 for evidence that the inoculation videos did not have the intended effect when participants were asked about sharing the content online). However, as in Experiment 1, there again were no significant effects on truth discernment. Surprisingly, there was no effect even when only analysing the subset of items where all false headlines were emotionally manipulative and all true headlines were not. Thus, even if one were to assume that the posited misinformation techniques are present only for misinformation (therefore presenting the most favourable context for testing the effect of the intervention), the inoculation apparently did not help participants spot misinformation.

To test this more directly, we reran the experiment but included only the high-emotionality false and low-emotionality true headlines. This was intended to rule out the possibility that including the fully crossed set of true/false and high/low-emotionality headlines undermined the effect of the intervention in some way.

https://doi.org/10.1038/s41562-024-02023-2

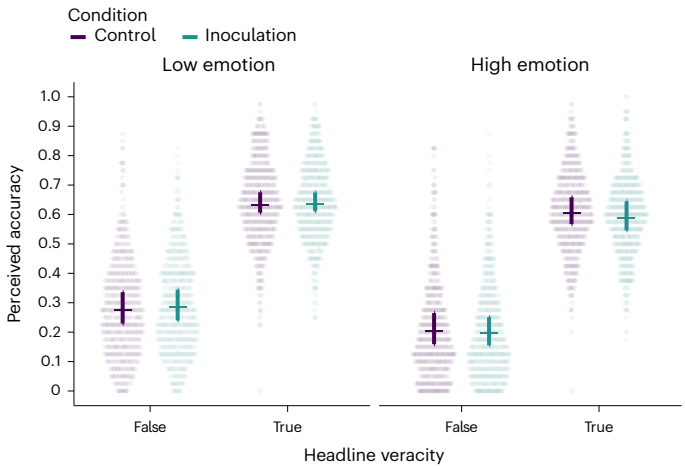

Fig. 2 | Perceived accuracy ratings from Experiment 2 by condition, headline veracity and emotional manipulativeness. Average adjusted predictions (horizontal lines) and 95% confidence intervals (vertical lines) are shown. Each dot is one participant's mean rating (n = 2,033). The predictions were obtained from OLS regression analysis with robust standard errors clustered on participant and item (two-sided tests without adjustments for multiple comparisons).

In Experiment 3, emotional manipulativeness and headline truth were manipulated to be fully consistent with each other. We thus ran a linear regression model with robust standard errors clustered on subject and headline, predicting accuracy judgements using an inoculation treatment dummy (0 = control, 1 = treatment), a news headline veracity/emotional manipulativeness dummy (0 = false/high emotion, 1 = true/low emotion) and their interaction (Supplementary Table 3). As expected, true low-emotion content was judged as more accurate than false high-emotion content in the control condition (b = 0.48 (0.42, 0.54), z = 15.48, P < 0.001). However, the inoculation video did not significantly increase the difference in perceived accuracy for true low-emotion versus false high-emotion content (b = −0.01 (−0.03, 0.02), z = −0.58, P = 0.559, BF = 0.009). That is, the participants did not improve in their ability to distinguish between true and false news headlines even though we manipulated the true headlines to be low emotion and the false headlines to be high emotion. There was no significant inoculation effect on accuracy judgements for either false high-emotion (b = 0.003 (−0.01, 0.02), z = 0.40, P = 0.691, d = −0.01, BF = 0.011) or true low-emotion (b = −0.004 (−0.02, 0.02), z = −0.37, P = 0.709, d = 0.01, BF = 0.011) headlines (Fig. 3).

Experiment 3 set a highly favourable context for testing the effect of the inoculation on discerning between true and false news headlines—that is, the participants only received false headlines that contained the emotional manipulation technique and true headlines that did not contain the technique; hence, identifying the technique should correspond directly to detecting misinformation. Nevertheless, the inoculation did not produce a significant difference between true (low emotion) and false (high emotion) videos. Thus, the emotional language inoculation (at least in the present form) appears to be very specific: it helps people identify emotionality in a general sense (as evidenced in Experiments 1 and 2), but its effect is undermined by some combination of (a) asking about accuracy (instead of the presence of the technique per se) and (b) the presence of claims that are either true or false (rather than the techniques presented on their own without claims of fact). That is, when the task is made more difficult by intermixing actual true or false claims, the video appears to lose its effectiveness as an inoculation against misinformation.

One potential reason for this is that the emotional language inoculation video does not actually draw attention to accuracy in any specific way. Rather, the video focuses on identifying emotionality techniques as a form of grabbing one's attention. Although the videos were proposed as a misinformation intervention, it may be that they do not actually draw people's attention to whether content is true or false. Other research has shown that, at least in terms of decisions about what to share online, users may often fail to even consider whether content is accurate before they decide to share that content[26]. Indeed, subtly reminding people about accuracy is sufficient to increase the quality of content (that is, less false relative to true content) that people are willing to share online[29]. One possibility, then, is that a simple reminder about the importance of considering accuracy (and the threat of misinformation) prior to the inoculation video may help users apply the specific content that they learn about emotional manipulation in the video. In support of this possibility, previous inoculation research that used longer videos with a broader range of techniques and that also included explicit reference to accuracy has found an effect on judgements of the reliability (a proxy for accuracy) of subsequent misinforming material[23] (albeit, as discussed, material that does not contain true or false claims per se). To test this possibility, for Experiment 4 we modified the inoculation video so that it was bookended with reminders about accuracy (the video can be viewed at https://osf.io/ym5wg/).

As in Experiment 3, emotional manipulativeness and headline truth were manipulated to be fully consistent with each other. We thus ran a linear regression model with robust standard errors clustered on subject and headline, predicting accuracy judgements using an inoculation (with accuracy prompts) treatment dummy (0 = control, 1 = treatment), a news headline veracity/emotional manipulativeness dummy (0 = false/high emotion, 1 = true/low emotion) and their interaction (Supplementary Table 4). True low-emotion content was judged as more accurate than false high-emotion content in the control condition (b = 0.45 (0.39, 0.51), z = 14.99, P < 0.001). Critically, the inoculation video did significantly increase the difference in perceived accuracy for true low-emotion versus false high-emotion content (b = 0.03 (0.01, 0.06), z = 2.95, P = 0.003, d = 0.16, 7.6% increase). That is, the combined inoculation and accuracy prompt video was successful in leading participants to improve in their ability to distinguish between true and false news headlines (when true headlines were manipulated to be low emotion and false headlines were manipulated to be high emotion). Interestingly, the inoculation did not have a significant effect on accuracy judgements for false high-emotion headlines (b = −0.01 (−0.03, 0.01), z = −1.08, P = 0.282, d = −0.036) but significantly increased accuracy judgements for true low-emotion headlines (b = 0.02 (0.01, 0.04), z = 2.79, P = 0.005, d = 0.089) (Fig. 4).

Experiment 4 provided evidence that the inoculation was effective at increasing truth discernment when paired with an accuracy prompt. Although this effect was small and evident only for true low-emotion headlines (as opposed to false high-emotion headlines), it nonetheless provides some evidence that participants were able to apply what they learned about emotional manipulation in the short video to content that contains true and false claims. Nonetheless, one question remains: was the accuracy prompt itself responsible for the increase in truth discernment, or are both elements necessary to produce the effect? On the basis of past theorizing[34], it is unlikely that the accuracy prompt alone would improve judgements of accuracy since the intervention simply draws attention to accuracy without teaching participants any new information. Furthermore, our participants were already being directly asked about accuracy as our central outcome measure—and thus the task is already directly drawing their attention to accuracy (asking about accuracy as an outcome has been shown to be a maximally strong accuracy prompt[26]). Nonetheless, we used a different accuracy prompt in this study than in past work, and we cannot rule out the possibility that it had some sort of carry-over effect on its own (for example, if it was doing more than just cueing accuracy, such as applying normative pressure or increasing motivations to be accurate). We therefore tested

https://doi.org/10.1038/s41562-024-02023-2

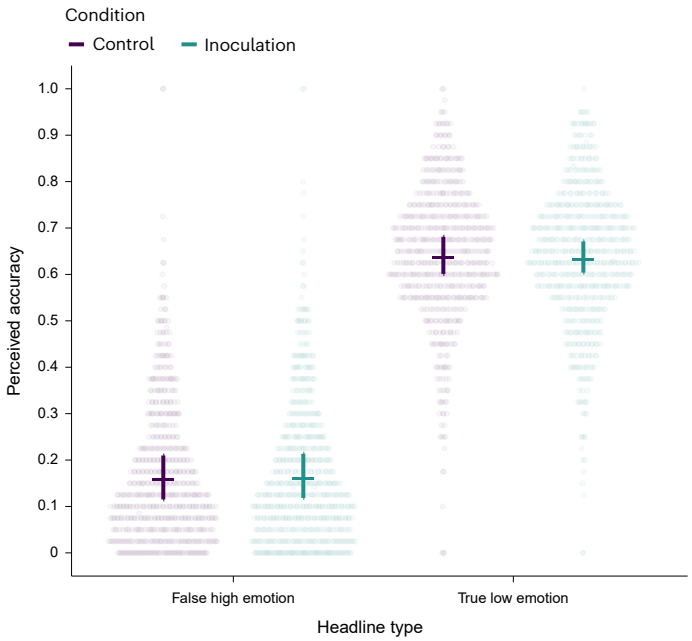

Fig. 3 | Perceived accuracy ratings from Experiment 3 by condition and headline type. Average adjusted predictions (horizontal lines) and 95% confidence intervals (vertical lines) are shown. Each dot is one participant's mean rating (n = 1,211). The predictions were obtained from OLS regression analysis with robust standard errors clustered on participant and item (two-sided tests without adjustments for multiple comparisons).

for carry-overs from the accuracy prompt in Experiment 5 by comparing the effect of the combined accuracy prompt and inoculation video with that of the accuracy prompt video on its own.

As preregistered, we first confirmed that there were no differences between the two sets of control videos (condition dummy: b = −0.01 (−0.03, 0.01), z = −0.90, P = 0.366, BF = 0.014; interaction between condition and veracity: b = 0.003 (−0.02, 0.03), z = 0.24, P = 0.807, BF = 0.007). We thus pooled the control conditions. We ran a linear regression model with robust standard errors clustered on subject and headline, predicting accuracy judgements using a combined inoculation and accuracy prompt treatment dummy (0 = control, 1 = treatment), an only accuracy prompt treatment dummy (0 = control, 1 = treatment), a news headline veracity/emotional manipulativeness dummy (0 = false/high emotion, 1 = true/low emotion), and the interactions between manipulativeness and our two treatment conditions (Supplementary Table 5). True low-emotion content was judged as more accurate than false high-emotion content in the control (b = 0.45 (0.39, 0.51), z = 14.34, P < 0.001).

Replicating Experiment 4, the inoculation video significantly increased the difference in perceived accuracy for true low-emotion versus false high-emotion content (b = 0.05 (0.02, 0.07), z = 3.40, P = 0.001, d = 0.22, 10.2% increase). In this case, the combined inoculation and accuracy prompt video did not significantly decrease perceptions of accuracy for false high-emotion headlines (b = −0.03 (−0.06, 0.00), z = −1.88, P = 0.060, d = −0.088) but did significantly increase accuracy judgements for true low-emotion headlines (b = 0.02 (0.00, 0.04), z = 2.14, P = 0.033, d = 0.090) (Fig. 5). As expected, there was no interaction effect for the accuracy prompt alone (b = 0.001 (−0.02, 0.02), z = 0.07, P = 0.948, BF = 0.009), indicating that simply drawing people's attention to the problem of misinformation is not sufficient to increase subsequent truth discernment.

Importantly, the effect of the combined treatment on discernment was significantly larger than the effect of the accuracy prompt alone (Wald test comparing (1) the coefficient on the interaction between accuracy prompt only and veracity with (2) the coefficient on the interaction between combined inoculation+accuracy and veracity, χ^2_1 = 11.33, P < 0.001). In other words, the accuracy prompt and inoculation treatment had a synergistic effect on truth discernment, but similar to prior experiments, neither was significantly effective on its own. Since prior work on accuracy prompts has focused on news headline sharing (and not belief, as we measure here), this research also indicates that inoculations may boost the effect of prompting accuracy beyond choices about what to share.

Discussion

A common sentiment is that multiple tools will be needed to effectively combat misinformation[2,35]. We present evidence here that popular intervention approaches can have synergistic effects, even when not effective on their own.

In particular, we tested a popular inoculation intervention intended to increase people's ability to recognize emotional language manipulation in text. Unlike past work, we tested this intervention using stimuli that not only contained the emotional language manipulation (in contrast to neutral/low-emotionality content) but also contained true or false claims—allowing us to assess the effect of the inoculation on truth discernment.

Our results identified a limitation of the emotional inoculation intervention. Consistent with past work[15], the intervention was successful in helping people distinguish content that contains (versus does not contain) an emotional language manipulation. Critically, however, we found consistent evidence that the inoculation on its own did not help people distinguish between true and false content. In fact, we did not find an inoculation effect on accuracy discernment even in the extreme case where all false claims—and no true claims—contained emotional manipulation.

Why did the emotional inoculation fail to facilitate truth discernment? One plausible mechanism is that the emotional inoculation that we focused on (which was among the strongest interventions in past work[15]) did not draw an explicit connection between the use of emotional language and falsehoods per se. This is important because past work shows that people tend to be inattentive to accuracy on social media[27]. Consistent with this, we found that focusing people's attention on accuracy before they watched the inoculation video (as is done in popular accuracy prompt interventions[29]) did lead to an increase in truth discernment as a consequence of the combined force of inoculation and accuracy prompting. This demonstrates a synergy between inoculation and accuracy prompt interventions, albeit with modest effect sizes (7.6–10.2% increase in accuracy discernment relative to the control, d = 0.16–0.22). Interestingly, however, this increase in truth discernment was primarily the result of increasing belief in true low-emotion claims (as opposed to decreasing belief in false high-emotion claims).

Implications

These findings emphasize the importance of jointly deploying multiple interventions. Building off recent theorizing about boost versus nudge interventions, our results suggest that an inoculation boost—that is, an intervention that increases people's cognitive and motivational competencies through learning[36]—is important for improving truth discernment capabilities, whereas an accuracy prompt nudge is critical for engaging/activating that capability.

Our findings have implications for work on psychological inoculation in the context of misinformation. In particular, misinformation inoculation interventions have typically been evaluated by testing their effects on people's ability to identify manipulative techniques. This is usually done using stylized or 'synthetic' materials that do not contain actual true or false claims, in an effort to offer 'clean' examples of the manipulative technique absent real-world truth claims that might influence people's evaluations. The logic behind this approach

https://doi.org/10.1038/s41562-024-02023-2

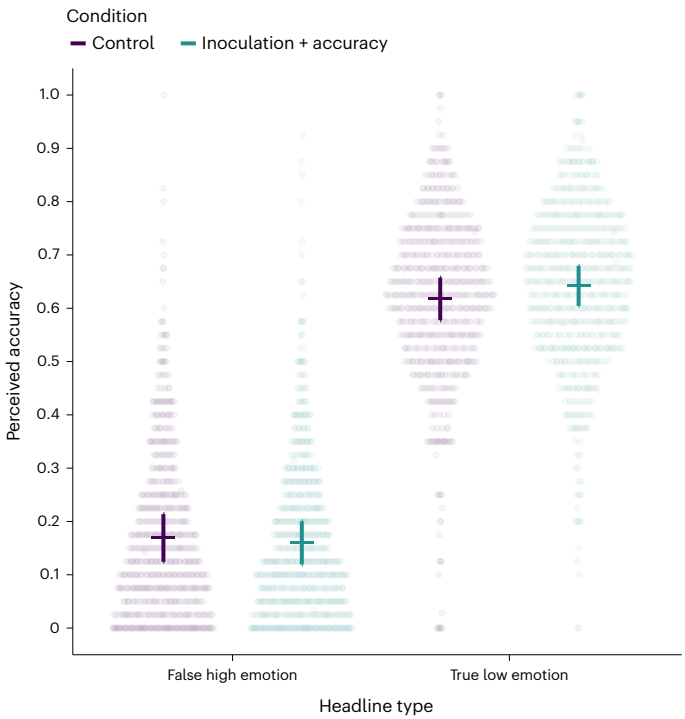

Fig. 4 | Perceived accuracy ratings from Experiment 4 by condition and headline type. Average adjusted predictions (horizontal lines) and 95% confidence intervals (vertical lines) are shown. Each dot is one participant's mean rating (n = 1,211). The predictions were obtained from OLS regression analysis with robust standard errors clustered on participant and item (two-sided tests without adjustments for multiple comparisons).

is that helping people recognize common manipulation techniques, such as emotional language manipulation, ought to help them distinguish a wide range of misinformation, which includes misleading content as well as false content. Our findings, however, indicate that simply providing information about manipulative techniques may not be sufficient to lead people to be 'inoculated' against belief in actual falsehoods. Rather, only when people were prompted to connect the information about emotional manipulation with factuality did we find evidence that the inoculation had an impact on truth discernment.

This implies that prior inoculation approaches may be broadly effective only if they have successfully connected the information that people are learning about manipulation techniques with the threat of misinformation. Moreover, even in such cases, past work that has tested inoculation approaches using technique identification may have created inflated expectations regarding effect sizes. For example, Experiment 1 found effect sizes of d = 0.41 and d = 0.58 on technique discernment using manipulativeness and trustworthiness ratings as the outcome (in line with Roozenbeek et al.'s[15] finding of d = 0.49 on technique discernment). In contrast, Experiment 2 found an effect size of only d = 0.20 on technique discernment when using identical materials but with perceived accuracy as the outcome—and, as mentioned above, it found no significant effect (d = −0.03) on truth discernment. Furthermore, when we did ultimately find a significant improvement in truth discernment from the combined inoculation+accuracy prompt intervention in Experiments 4 and 5, this effect was driven more by increased belief in low-emotion true headlines than by decreased belief in high-emotion false headlines (that is, the effect was significant for low-emotion true headlines in both studies, but non-significant for high-emotion false in Experiment 4 and failed to reach significance in Experiment 5, P = 0.06). The participants were taught that emotionality should be viewed with scepticism, and this led to an increase in perceived accuracy for true content that did not contain emotional language. This is similar to the 'implied truth effect', wherein putting warning labels on some news may increase belief in unlabelled news[37]. Nonetheless, it is surprising that the inoculation did not have a stronger negative effect on the high-emotion false claims than the positive effect on low-emotion true claims. Future work should reconcile this pattern with inoculation theory's focus on reducing susceptibility to manipulative content per se.

Limitations

One important empirical limitation of our research is that we found a successful inoculation (with accuracy prompt) effect on truth discernment in a context where the emotional language manipulation was fully confounded with truth. That is, the false content all contained an emotional language manipulation, and the true content did not contain any emotional language manipulation. There is reason to expect false content to be somewhat more likely to contain emotional manipulation than true content. For example, Carrasco-Farré[38] showed in an analysis of nearly 100,000 news articles that fake news articles contain more negative sentiment on average than factual articles. However, emotional language and truth/falsity are certainly far from perfectly correlated in the real world. Thus, to whatever extent a disconnect exists between emotionality and veracity outside the lab setting, the true effect on truth discernment will be smaller than the average effect of d = 0.20 observed for the combined intervention in our studies.

Related to this, although our news headlines were sourced from real-world content, our emotional language manipulation was artificial (in the sense that we created the language of the manipulation ourselves). Although this is typical of past work (for example, ref. 15), it is possible that our manipulations were weaker (or stronger) than would be common in real-world content. Our pretest allowed us to validate that our emotional language manipulation successfully induced emotions (a step not common in past work), but this does not necessarily guard against the possibility that our manipulation was unrepresentative of emotional content in the field.

Another limitation of the present studies is the reliance on headlines alone, which provides a very limited attack surface for the inoculation intervention to find traction. It is therefore possible that a more realistic extent of content (for example, a headline plus a lede or even an entire news article) will provide more opportunity for the intervention to 'grab'. Although this remains an open empirical question, it is quite plausible that inoculation effect sizes may be greater when there is more content that the intervention can be applied to.

Furthermore, this study tested only one medium and format of interventions for both inoculation and accuracy prompts—animated video. It is worth testing variations of videos—live action, creator driven, humorous—as well as other mediums such as text, audio or captioned images. Similarly, we used a specific set of 32 headlines. Future work should investigate how the results generalize to other headline sets. We also focused only on emotional language manipulation. Although this is a particularly salient and common misinformation technique[38,39], it is nonetheless important to investigate whether our results generalize to other misinformation techniques. This research indicates that inoculations from past and present research—which tend to be quite variable—may help accuracy discernment if they contain messages that specifically prompt people to translate what they are learning to truth/falsity[40].

Future work should look into the mechanism at play in the accuracy prompt used in our study. It is unclear whether the prompt in our study was priming accuracy (as in past work on sharing intentions) or something else, such as linking the inoculation to misinformation. The portion of the video including the prompt did mention accuracy, but accuracy was not the primary focus. Past work has used a variety of accuracy prompt methods[28,29]; however, one limitation is that these tend to be varied and may differ in terms of how closely connected they

https://doi.org/10.1038/s41562-024-02023-2

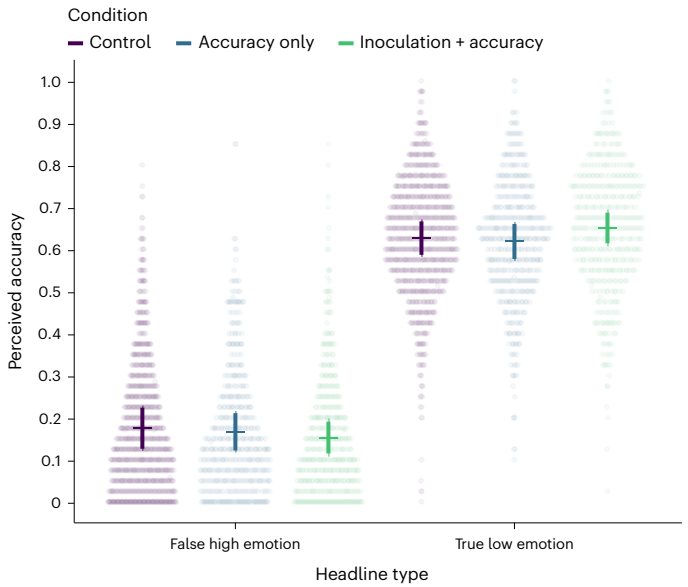

Fig. 5 | Perceived accuracy ratings from Experiment 5 by condition and headline type. Average adjusted predictions (horizontal lines) and 95% confidence intervals (vertical lines) are shown. Each dot is one participant's mean rating (n = 1,804). The predictions were obtained from OLS regression analysis with robust standard errors clustered on participant and item (two-sided tests without adjustments for multiple comparisons).

are to the key mechanism(s) of interest. Future work clarifying this issue would be beneficial.

Conclusion

We conclude that combining interventions can have synergies for helping people better distinguish between true and false information online. Theoretical modelling has suggested the existence of synergies from combining approaches to fight misinformation[2,35]. The present studies present empirical evidence of this phenomenon using randomized-controlled trials with participants viewing scalable interventions and real-world headlines. Our study has significant implications for the burgeoning field of designing misinformation interventions. The practitioners, technologists and policymakers developing these interventions would benefit from testing and deploying multiple methods to fight misinformation in tandem where they are synergistic. Combined interventions can remain highly scalable; in this case, both interventions were delivered together in a short video (about two minutes).

Notwithstanding the potential success of such scalable interventions, they can only be part of the solution to the misinformation problem. Misinformation does not arise in a vacuum but flourishes as part of an online ecosystem whose architecture focuses on capturing user attention[4]. Given that purveyors of misinformation are, by definition, freed from constraints imposed by reality, they can exploit known biases in human attention (for example, by presenting outraging content[41], which platform algorithms are likely to amplify). Interventions such as those developed here must therefore be accompanied by creating an Internet with democratic credentials and based on user empowerment[42].

Methods

This research complied with all ethical regulations of and was approved by the Research Ethics Committee at the University of Regina (no. 2018-116). Informed consent was obtained from all participants (including those in the pretest), and they were compensated on the basis of hourly rates standard to the platforms from which they were recruited. No statistical methods were used to predetermine sample sizes. Nonetheless, for each study we performed a sensitivity analysis through simulations to estimate the smallest effect size (Cohen's d) we could detect with at least 80% statistical power, given the mixed-factorial design and sample size. Across the five studies, the smallest effect size we could detect was d = ~0.10 (d for the five studies was 0.12, 0.13, 0.11, 0.11 and 0.10). In all experiments, the participants were randomized into conditions (maintaining blind assignment).

Experiment 1

Participants. Participants were recruited from Amazon's Mechanical Turk via CloudResearch, which has been shown to provide high-quality participants[43]. The participants were randomized into an emotional inoculation treatment condition or a control condition, and those in the control condition were further randomized to one of three control videos. In total, 1,268 participants entered the experiment. However, 33 participants indicated not having social media and therefore did not qualify for the study, and 6 participants failed an attention check at the beginning of the study and did not continue. In addition, 3 participants were removed from the survey because they indicated not being able to turn on their volume (which was critical for watching the videos), and 18 participants quit the survey prior to the intervention. A further 178 participants dropped out and did not contribute data to the primary task. The fraction of participants dropping out did not vary significantly across conditions (15.2% in control, 14.2% in treatment; $\chi^2_1 = 0.23$, $P = 0.630$). We therefore had a final sample size of 1,030 participants (mean age, 36 years), of whom 364 were male, 610 were female, 30 chose some other response (for example, trans/non-binary) and 26 did not respond to the gender question.

Materials. We used 16 true and 16 false headlines to create 32 emotionally charged (fear or anger evoking) variants and 32 emotionally neutral variants by manipulating the words in the headlines. We also used an emotion inoculation video for the treatment condition and an educational video on curling, bananas or freezer burn for the control condition. The full materials and data can be found at https://osf.io/ym5wg/; the inoculation video is available at https://inoculation.science/.

Procedure. The participants completed the first attention check in the beginning of the study and then proceeded to provide their consent, which was followed by the second attention check question. The second attention check question was an eliminating question, and participants who failed this check were redirected to the end of the study. The participants then answered two questions on their social media usage: the type of content they would consider sharing on social media (such as political news, sports news and business news) and the social media platforms that they use (such as Facebook, Twitter and Instagram). Participants who indicated that they did not use any social media platform were redirected to the end of the study as well. Since the study entailed watching a video, we next asked the participants if the volume on their device was on. For participants who answered in the negative, we provided an additional opportunity to turn the volume on. Those who could not or indicated being unwilling to do so were redirected to the end of the study.

The participants were then randomly assigned to either the treatment or the control condition. Participants in the treatment condition viewed an emotion inoculation video, whereas participants in the control condition viewed an educational video of a similar length on one of three topics: curling, bananas or freezer burn. The videos were set up such that they would autoplay and were not interactable. Therefore, the participants could not pause the videos once they automatically started playing. Moreover, the next button was not made visible till the video was done playing.

The participants then read the instructions about viewing and rating 32 social media headlines. A total of 64 social media headlines

generated for this study were evenly split between two blocks. The first block had 16 false and 16 true headlines, with 8 emotionally charged and 8 emotionally neutral headlines for each veracity. The second block had emotionally opposite counterparts of the headlines in the first block. Each participant saw only one version (high or low emotionality) of each headline, and this presentation was counterbalanced across participants. After viewing the randomly selected headlines, the participants rated the trustworthiness and the manipulativeness of each headline on a six-point scale: (1) extremely untrustworthy to (6) extremely trustworthy, and (1) not at all manipulative to (6) extremely manipulative.

All participants completed the third attention check question after rating the headlines. The participants then completed two measures outside the scope of the present investigation—the actively open-minded thinking scale[44] and the Emotion Regulation Questionnaire[45]. The participants then completed the last attention check question, followed by questions related to their political position and their political preference in general on social issues and economic issues. Finally, the participants answered demographics questions, which entailed measurement of age, gender, education, income and ethnicity. The participants also indicated how much they believe in God(s). These measures are also outside the scope of the present investigation, and we do not report the results.

We note that the first version of this experiment was run using participants recruited from Lucid[46]. In the Lucid experiment, the participants were asked about manipulativeness and trustworthiness (as here) or about accuracy (as in Experiment 2) or about social media sharing intentions. However, there were issues with inattentiveness (30.7% of participants who passed the initial A/V check failed at least one trivial attention check in the Lucid experiment, compared with only 5.6% in the Mechanical Turk experiment), and significantly more participants dropped out during the inoculation video compared with the control videos ($\chi^2_1 = 5.26$, $P = 0.023$), which violates random assignment and undermines causal inference. We thus do not consider the results of the Lucid experiment to provide useful insight (although for completeness, we report them in Supplementary Section 2).

Our preregistration (https://osf.io/ym5wg/) included political salience as a factor, but we report the analysis without this factor included because it is not central to the research question, and including it leads to an overly complicated model with four-way interactions.

Experiment 2

Participants. Participants were again recruited from Amazon's Mechanical Turk via CloudResearch. In total, 2,471 participants entered the experiment and were randomly assigned to rate the accuracy of the headlines or to indicate whether they would share them online. However, 59 participants indicated not having social media and therefore did not qualify for the study, and 10 participants failed an attention check at the beginning of the study and did not continue. In addition, 5 participants were removed from the survey because they indicated not being able to turn on their volume, and 47 participants quit the survey prior to the intervention. A further 317 participants dropped out and did not contribute data to the primary task. The fraction of participants dropping out did not vary significantly across conditions (14.3% in control, 12.6% in treatment; $\chi^2_1 = 1.50$, $P = 0.221$). We therefore had a final sample size of 2,033 participants (mean age, 40 years), of whom 793 were male, 1,163 were female, 36 chose some other response (for example, trans/non-binary) and 41 did not respond to the gender question.

Materials. The materials were identical to those in Experiment 1.

Procedure. Unlike Experiment 1, this study had a 2 (condition: treatment versus control) × 2 (question asked: accuracy versus sharing) design. Instead of rating trustworthiness and manipulativeness, the participants were randomly assigned to either rate the accuracy of the headline (using a scale of 1 = extremely inaccurate to 6 = extremely accurate) or indicate how likely they would be to share the headline on social media (using a scale of 1 = extremely unlikely to 6 = extremely likely). The procedure of this study was identical to that of Experiment 1 in all other aspects.

We focus here on the results for the accuracy condition. The results for the sharing condition are presented in Supplementary Section 3. In short, the inoculation did not increase either emotion or truth discernment for sharing. Note that the preregistration for Experiment 1 mentions the Experiment 2 dependent variables. We ran Experiment 1 12 days prior to Experiment 2 (to ensure successful technique recognition) but preregistered them together as one experiment with three dependent variables. We report them separately here for simplicity.

Experiment 3

Participants. US participants were recruited from Prolific Academic. In total, 1,526 participants entered the experiment and were randomly assigned to the treatment or the control. However, 48 participants indicated not having social media and therefore did not qualify for the study, and 7 participants failed an attention check at the beginning of the study and did not continue. No participants were removed from the survey because they indicated not being able to turn on their volume, but 9 participants quit the survey prior to the intervention. A further 254 participants dropped out and did not contribute data to the primary task. The fraction of participants dropping out did not vary significantly across conditions (17.2% in control, 17.6% in treatment; $\chi^2_1 = 0.04$, $P = 0.841$). We therefore had a final sample size of 1,208 participants (mean age, 40 years), of whom 652 were male, 517 were female, 28 chose some other response (for example, trans/non-binary) and 11 did not respond to the gender question.

Materials and procedure. Of the 32 emotionally charged (16 true and 16 false) and 32 emotionally neutral (16 true and 16 false) headlines used in Experiment 1, we kept only the 16 false emotionally charged headlines and the 16 true emotionally neutral headlines for use in this study.

The participants provided accuracy ratings as in Experiment 2 (there was no sharing condition). The 16 headlines seen by the participants came from one of two randomly chosen blocks. The first block had eight false emotionally charged and eight true emotionally neutral headlines. The second block had a different set of false emotionally charged and true emotionally neutral headlines, eight of each kind. The remaining procedure was identical to that of Experiment 2, except that we removed the actively open-minded thinking scale, the Emotion Regulation Questionnaire and the fourth attention check (owing to the shorter length). The analyses were preregistered (https://osf.io/ym5wg/).

Note that we ran a parallel experiment using participants from Mechanical Turk ($n = 1{,}023$) that is reported in Supplementary Section 4. We similarly did not find an effect of the emotional inoculation ($b = -0.003$, $P = 0.734$); however, we believed that the data were questionable because the difference between true and false headlines, which is robustly observed to be large in all other studies in this paper as well as in the literature more generally, was extremely small ($b = 0.04$, whereas $b = 0.48$ in Experiment 3). We thus do not report this study in the main text.

Experiment 4

Participants. US participants were recruited from Prolific Academic. In total, 1,348 participants entered the experiment and were randomly assigned to the treatment or the control. However, 27 participants indicated not having social media and therefore did not qualify for the study, and 4 participants failed an attention check at the beginning of the study and did not continue. No participants were removed from the survey because they indicated not being able to turn on their volume or failed to reach the intervention. A further 95 participants dropped out

and did not contribute data to the primary task. The fraction of participants dropping out did not vary significantly across conditions (6.8% in control, 7.8% in treatment; $\chi^2_1 = 0.45$, $P = 0.504$). We therefore had a final sample size of 1,211 participants (mean age, 40 years), of whom 647 were male, 514 were female, 37 chose some other response (for example, trans/non-binary) and 13 did not respond to the gender question.

Materials and procedure. The video used in this study was modified to include accuracy prompts. This new video had accuracy reminders in the beginning and at the end, with the inoculation video used in the previous experiments in the middle (https://osf.io/ym5wg/). The goal was to increase the chances that the viewer would apply the content of the video to the issue of truth versus falsity. The three educational videos in the control condition were also replaced to match the duration of the new treatment video. The new educational control videos were about stars, Scotland and bats. The headlines and procedure were the same as in Experiment 3. The analyses were preregistered (https://osf.io/ym5wg/).

Experiment 5

Participants. US participants were recruited from Prolific Academic. In total, 1,996 participants entered the experiment and were randomly assigned to the treatment or the control. However, 62 participants indicated not having social media and therefore did not qualify for the study, and 4 participants failed an attention check at the beginning of the study and did not continue. In addition, 1 participant was removed from the survey because they indicated not being able to turn on their volume, and 11 participants quit the survey prior to the volume question. A further 114 participants dropped out and did not contribute data to the primary task. The fraction of participants dropping out did not vary significantly across conditions (6.0% in long control, 6.6% in combined inoculation and accuracy, 5.5% in short control, 5.8% in accuracy only; $\chi^2_3 = 0.59$, $P = 0.899$). We therefore had a final sample size of 1,804 participants (mean age, 40 years), of whom 906 were male, 838 were female, 49 chose some other response (for example, trans/non-binary) and 11 did not respond to the gender question.

Materials and procedure. In this study, we had two sets of videos. As in Experiment 4, we had an inoculation video with accuracy reminders in the beginning and at the end for the treatment condition, and three duration-matched educational videos (on stars, Scotland and bats) for the control condition. Unlike Experiment 4, we additionally created a condition where participants only received the initial accuracy-prompt component of the video. We therefore used a different set of three duration-matched educational/motivational videos (on recycling, entrepreneurship and effort) for the control condition. The materials and procedure were otherwise identical to those of Experiment 4. The analyses were preregistered (https://osf.io/ym5wg/).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The materials, data and preregistrations for these experiments can be found at https://osf.io/ym5wg/.

Code availability

The code for these experiments can be found at https://osf.io/ym5wg/.

References

1. Athey, S., Grabarz, K., Luca, M. & Wernerfelt, N. Digital public health interventions at scale: the impact of social media advertising on beliefs and outcomes related to COVID vaccines. *Proc. Natl Acad. Sci. USA* **120**, e2208110120 (2023).
2. Bak-Coleman, J. B. et al. Combining interventions to reduce the spread of viral misinformation. *Nat. Hum. Behav.* **6**, 1372–1380 (2022).
3. Ecker, U. K. H. et al. The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* **1**, 13–29 (2022).
4. Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the Internet: confronting digital challenges with cognitive tools. *Psychol. Sci. Public Interest* **21**, 103–156 (2020).
5. Kozyreva, A. et al. Toolbox of individual-level interventions against online misinformation. *Nat. Hum. Behav.* **8**, 1044–1052 (2024).
6. Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends Cogn. Sci.* https://doi.org/10.1016/j.tics.2021.02.007 (2021).
7. Chan, M. P. S., Jones, C. R., Hall Jamieson, K. & Albarracín, D. Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. *Psychol. Sci.* **28**, 1531–1546 (2017).
8. Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N. & Cook, J. Misinformation and its correction: continued influence and successful debiasing. *Psychol. Sci. Public Interest* **13**, 106–131 (2012).
9. Nieminen, S. & Rapeli, L. Fighting misperceptions and doubting journalists' objectivity: a review of fact-checking literature. *Polit. Stud. Rev.* **17**, 296–309 (2019).
10. 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. USA* **118**, e2104235118 (2021).
11. Wood, T. & Porter, E. The elusive backfire effect: mass attitudes' steadfast factual adherence. *Polit. Behav.* **41**, 135–163 (2019).
12. Cook, J. et al. Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence. *PLoS ONE* **12**, e0175799 (2017).
13. Lewandowsky, S. & van der Linden, S. Countering misinformation and fake news through inoculation and prebunking. *Eur. Rev. Soc. Psychol.* **32**, 348–384 (2021).
14. Traberg, C. S., Roozenbeek, J. & van der Linden, S. Psychological inoculation against misinformation: current evidence and future directions. *Ann. Am. Acad. Polit. Soc. Sci.* **700**, 136–151 (2022).
15. 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).
16. Roozenbeek, J. & van der Linden, S. How to combat health misinformation: a psychological approach. *Am. J. Health Promot.* **36**, 569–575 (2022).
17. Compton, J. in *The SAGE Handbook of Persuasion: Developments in Theory and Practice* (eds Dillard, J. P. & Shen, L.) 220–236 (SAGE, 2013).
18. Compton, J., van der Linden, S., Cook, J. & Basol, M. Inoculation theory in the post-truth era: extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Soc. Pers. Psychol. Compass* **15**, e12602 (2021).
19. Jigsaw. Defanging disinformation's threat to Ukrainian refugees. *Medium* https://medium.com/jigsaw/defanging-disinformations-threat-to-ukrainian-refugees-b164dbbc1c60 (2023).
20. Klepper, D. Google to expand misinformation 'prebunking' in Europe. *AP News* https://apnews.com/article/technology-science-politics-germany-business-a10273eea5a0227c38187cc4f84d8788 (2023).
21. Mukherjee, S. & Coulter, M. Exclusive: Google launches anti-misinformation campaign in India. *Reuters* https://www.reuters.com/technology/google-launches-anti-misinformation-campaign-india-2022-12-06/ (6 December 2022).

22. Traberg, C. S. et al. in *Managing Infodemics in the 21st Century: Addressing New Public Health Challenges in the Information Ecosystem* (eds Purnat, T. D. et al.) 99–111 (Springer International, 2023); https://doi.org/10.1007/978-3-031-27789-4_8
23. Lewandowsky, S. & Yesilada, M. Inoculating against the spread of Islamophobic and radical-Islamist disinformation. *Cogn. Res. Princ. Implic.* **6**, 57 (2021).
24. Arechar, A. A. et al. Understanding and combatting misinformation across 16 countries on six continents. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-023-01641-6 (2023).
25. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D. G. Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy nudge intervention. *Psychol. Sci.* **31**, 770–780 (2020).
26. Pennycook, G. et al. Shifting attention to accuracy can reduce misinformation online. *Nature* https://doi.org/10.1038/s41586-021-03344-2 (2021).
27. Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G. & Rand, D. The social media context interferes with truth discernment. *Sci. Adv.* **9**, eabo6169 (2023).
28. Epstein, Z. et al. Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-71 (2021).
29. Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022).
30. Lin, H., Pennycook, G. & Rand, D. G. Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition* **230**, 105312 (2023).
31. Dias, N., Pennycook, G. & Rand, D. G. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-001 (2020).
32. Guay, B., Berinsky, A. J., Pennycook, G. & Rand, D. How to think about whether misinformation interventions work. *Nat. Hum. Behav.* **7**, 1231–1233 (2023).
33. Modirrousta-Galian, A. & Higham, P. A. Gamified inoculation interventions do not improve discrimination between true and fake news: reanalyzing existing research with receiver operating characteristic analysis. *J. Exp. Psychol. Gen.* **152**, 2411–2437 (2023).
34. Pennycook, G. & Rand, D. G. Nudging social media toward accuracy. *Ann. Am. Acad. Polit. Soc. Sci.* https://doi.org/10.1177/00027162221092342 (2022).
35. Bode, L. & Vraga, E. The Swiss cheese model for mitigating online misinformation. *Bull. At. Sci.* **77**, 129–133 (2021).
36. Hertwig, R. & Grüne-Yanoff, T. Nudging and boosting: steering or empowering good decisions. *Perspect. Psychol. Sci.* **12**, 973–986 (2017).
37. Pennycook, G., Bear, A., Collins, E. & Rand, D. G. The implied truth effect: attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *Manage. Sci.* **66**, 4921–5484 (2020).
38. Carrasco-Farré, C. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanit. Soc. Sci. Commun.* **9**, 162 (2022).
39. Paschen, J. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *J. Prod. Brand Manage.* **29**, 223–233 (2019).
40. Fazio, L. et al. Combating misinformation: a megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/uyjha (2024).
41. Bakir, V. & McStay, A. Fake news and the economy of emotions. *Digit. Journal.* **6**, 154–175 (2018).
42. Lewandowsky, S. & Pomerantsev, P. Technology and democracy: a paradox wrapped in a contradiction inside an irony. *Mem. Mind Media* **1**, e5 (2022).
43. Douglas, B. D., Ewell, P. J. & Brauer, M. Data quality in online human-subjects research: comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* **18**, e0279720 (2023).
44. Newton, C., Feeney, J. & Pennycook, G. On the disposition to think analytically: four distinct intuitive-analytic thinking styles. *Pers. Soc. Psychol. Bull.* https://doi.org/10.1177/01461672231154886 (2023).
45. Spaapen, D. L., Waters, F., Brummer, L., Stopa, L. & Bucks, R. S. The Emotion Regulation Questionnaire: validation of the ERQ-9 in two community samples. *Psychol. Assess.* **26**, 46–54 (2014).
46. Coppock, A. & Mcclellan, O. A. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res. Polit.* https://doi.org/10.1177/2053168018822174 (2019).

Acknowledgements

This research was supported by funding from Jigsaw (Google). G.P. acknowledges financial support from the Social Sciences and Humanities Research Council of Canada (grant no. 435-0806-2020) and the John Templeton Foundation (grant no. 61779). D.G.R. and A.J.B. acknowledge support from the National Science Foundation (NSF Award No. 2047152) and the Alfred P. Sloan Foundation (grant no. 2021-16891). A.J.B. also acknowledges support from the Russell Sage Foundation (grant no. 034076-0001). S.L. acknowledges financial support from the European Research Council (ERC Advanced Grant No. 101020961 PRODEMINFO), the Humboldt Foundation through a research award, the Volkswagen Foundation (grant 'Reclaiming individual autonomy and democratic discourse online: how to rebalance human and algorithmic decision making') and the European Commission (Horizon 2020, grant no. 101094752 SoMe4Dem). S.L. also receives funding from UK Research and Innovation (through EU Horizon replacement, grant no. 10049415). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding bodies. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

G.P. and D.G.R. designed the studies, analysed the data and wrote the paper. H.L. also contributed to data analysis and data visualization. G.P. and P.B. collected the data and designed/tested the stimuli. A.J.B., S.L., B.G. and R.C. contributed to the design of the studies and commented and edited drafts of the paper.

Competing interests

G.P. and D.G.R. have received funding from Meta and Google; G.P., A.J.B. and R.C. were employed at Google when the studies were designed and the data were collected; and B.G. both was and is employed at Google. Google has invested in both inoculation and accuracy prompts as misinformation interventions (in addition to supplying funding for these studies). The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-02023-2.

Correspondence and requests for materials should be addressed to Gordon Pennycook.

Peer review information *Nature Human Behaviour* thanks Lauri Rapeli and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Gordon Pennycook

Last updated by author(s): Sep 27, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Qualtrics survey software was used for data collection.
Data analysis	Stata (15) and R (4.3.1; 2023-06-16) were used for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Full materials and data can be found on OSF (https://osf.io/ym5wg/?view_only=c7bac9dea71143c2b8fcced7422f661f)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	In each study, we ask participants to indicate their gender (participants are free to skip the question). This is reported for every study. Consent was obtained.
Reporting on race, ethnicity, or other socially relevant groupings	Participants indicated their race in our surveys, but we do not report the racial breakdown in the manuscript.
Population characteristics	Average age was ~40 years old across our experiments. We had more individuals who identified as woman than men in our studies. The absolute number varies from study to study and this is reported in the main text.
Recruitment	Participants were recruited from online sources for survey research.
Ethics oversight	These studies were approved by the University of Regina Research Ethics Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Participants were given either an "inoculation" intervention or a control video and asked to assess a set of news headlines that varied in terms of emotional manipulativeness and truth/falsity.
Research sample	Participants were recruited from online panel services: Mechanical Turk, Prolific, and Lucid. Each of these sources has been validated in past work for research in behavioral science. Neither sample is truly representative, but Lucid uses quota-matching to achieve some level of representativeness. Mechanical Turk and Prolific are not representative (despite being quite diverse).
Sampling strategy	Participants were recruited online. We chose large samples and capped sample sizes based on the amount of money that we were willing to pay for each study.
Data collection	Participants completed surveys using Qualtrics software. These studies were all online and researchers were not physically present. The experiments are therefore fully blinded.
Timing	Data collection began on April 14th, 2022 and continued intermittently until January 19th, 2023.
Data exclusions	No participants with complete data were excluded
Non-participation	E1: 178, E2: 317, E3: 254, E4: 95, E5: 114
Randomization	Participants were randomized into conditions

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging