



# Kent Academic Repository

**Biddlestone, Mikey, Roozenbeek, Jon, Suiter, Jane, Culloty, Eileen and van der Linden, Sander (2025) *Tune in to the prebunking network! Development and validation of six inoculation videos that prebunk manipulation tactics and logical fallacies in misinformation*. Political Psychology, 46 (6). pp. 1858-1886. ISSN 1467-9221.**

## Downloaded from

<https://kar.kent.ac.uk/109539/> The University of Kent's Academic Repository KAR

## The version of record is available from

<https://doi.org/10.1111/pops.70015>

## This document version

Publisher pdf

## DOI for this version

## Licence for this version

CC BY (Attribution)

## Additional information

## Versions of research works

### Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

### Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal**, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

## Enquiries

If you have questions about this document contact [ResearchSupport@kent.ac.uk](mailto:ResearchSupport@kent.ac.uk). Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

# Tune in to the prebunking network! Development and validation of six inoculation videos that prebunk manipulation tactics and logical fallacies in misinformation

Mikey Biddlestone<sup>1</sup> | Jon Roozenbeek<sup>2</sup> | Jane Suiter<sup>3</sup> |  
Eileen Culloty<sup>3</sup> | Sander van der Linden<sup>4</sup>

<sup>1</sup>School of Psychology, University of Kent, Canterbury, UK

<sup>2</sup>Department of War Studies, King's College London, London, UK

<sup>3</sup>School of Communications, Dublin City University, Dublin, Ireland

<sup>4</sup>Department of Psychology, University of Cambridge, Cambridge, UK

## Correspondence

Mikey Biddlestone, School of Psychology, University of Kent, Canterbury, UK.  
Email: [mikeybiddlestone@gmail.com](mailto:mikeybiddlestone@gmail.com)

## Funding information

IRIS Infodemic, Grant/Award Number: SCH-00001-3391

## Abstract

Meta-analyses have demonstrated how inoculation interventions increase the detection of misinformation, but their scalability has remained elusive. To address this, Study 1 (pre-registered;  $N=1,583$ ) tested the efficacy of three short inoculation videos (prebunks) against three common manipulation tactics used in misinformation: (1) polarization, (2) conspiracy theories, and (3) fake experts. Results indicated that all three inoculation videos (vs. control) increased the detection of relevant manipulative content without altering perceptions of non-manipulative content, but only the polarization inoculation video increased manipulation discernment (i.e., increased ability to distinguish between manipulative and non-manipulative content). In Study 2 (pre-registered;  $N=1,603$ ), we tested the efficacy of three more inoculation videos containing logic-based prebunks against logical fallacies commonly used in misinformation: (1) whataboutism, (2) the moving the goalposts fallacy, and (3) the strawman fallacy. Detection of the relevant fallacious content was higher in all conditions (vs. control), but only the strawman fallacy inoculation video increased fallacy discernment. The moving the goalposts fallacy inoculation video appeared to increase overall distrust of relevant content, whereas the other two videos did not alter perceptions of relevant non-fallacious content. We discuss the implications and limitations of these findings.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Political Psychology* published by Wiley Periodicals LLC on behalf of International Society of Political Psychology.

## KEYWORDS

inoculation, logical fallacies, manipulation, misinformation, prebunking

## INTRODUCTION

Ranked as a top societal risk by the World Economic Forum (2024), the global spread of misinformation poses a threat to science, public health, and the integrity of democracies and elections worldwide (Lewandowsky et al., 2023; van der Linden & Roozenbeek, 2024). A recent consensus report from the American Psychological Association (APA) defines misinformation as “any information that is demonstrably false or otherwise misleading, regardless of its source or intent” (van der Linden et al., 2023, p. 7). Because outright falsehoods are relatively less common than content that is generally misleading (Allen et al., 2020), the bigger problem is content which involves some grain of truth alongside a known manipulation technique. Manipulation techniques can be defined as the strategic use of communication to alter people's beliefs, perceptions, or attitudes that tend to detract from the lack of evidential support for a claim, usually to achieve some agenda (see Lewandowsky et al., 2017; Roozenbeek & van der Linden, 2019). Thus, logical fallacies—flawed reasoning that fails to adhere to the principles of logic (see Tindale, 2007; Walton, 1995)—can also be used manipulatively to achieve the same outcomes.

One troubling example of the manipulative use of logical fallacies can be found in a 2021 headline from the *Chicago Tribune*; “A healthy doctor died two weeks after getting a COVID-19 vaccine” went viral and created widespread misperceptions because it falsely implied causation where there was only correlation (a known logical fallacy, see van der Linden & Roozenbeek, 2024; Kuru, 2024). By combining controlled experimental trials with exposure data from social media, recent research estimates that misleading headlines such as these are 46 times more damaging to vaccine intentions than outright fact-checked falsehoods (Allen et al., 2024; van der Linden & Kyrychenko, 2024). Misleading information can also undermine citizens' shared trust in democratic institutions (see Lewandowsky, 2024; Lewandowsky et al., 2024). This has led to policy recommendations for the promotion of media literacy education around support for democracy and civic values in the hopes of protecting citizens against the eroding impact of misinformation on democratic functioning (e.g., Lessenski, 2023).

One way this can be measured is through improving the detection of manipulation, and another is through manipulation discernment—the ability to distinguish manipulative information from non-manipulative information. Meta-analytic evidence (Lu et al., 2023) has demonstrated the efficacy of certain interventions to increase the credibility discernment of information, namely *psychological inoculation*: the motivation of psychological defense against manipulation, which involves preemptive refutations of how this manipulation might occur, and the presentation of *defanged microdoses* of misinformation to test subsequent *psychological immunity* against this manipulation (Ivanov et al., 2023; Lu et al., 2023; Roozenbeek et al., 2022).

## Inoculation against misinformation

Psychological inoculation is a theory of resistance to persuasion developed by McGuire in the 1960s (see McGuire, 1961). By preemptively refuting a weakened dose of persuasive tactics to participants, McGuire and Papageorgis (1961) demonstrated that preventing persuasion was particularly effective at conferring psychological resistance. This work has been compared to a *psychological vaccine* (see van der Linden, 2023), wherein participants are psychologically

immunized against the virus of persuasion by misinformation through the administration of weakened doses of the tactics used to persuade or manipulate, encouraging the production of psychological antibodies (e.g., increased detection, increased motivation and counterarguing) against these manipulation attempts.

Recent psychological research has seen a boom in the use of inoculation to reduce susceptibility to misinformation. The efficacy of this approach is frequently demonstrated through technique-based inoculation texts that preemptively refute the tactics that those spreading disinformation may use to manipulate. For example, Cook et al. (2017) found that prebunking the flawed argumentation techniques of false-balance (presenting niche contrarian views on scientific subjects) and giving a false sense of debate (distorting the fact that there is an overwhelming scientific consensus on a subject) reduced participants' susceptibility to climate change misinformation (see also van der Linden et al., 2017).

Another inoculation strategy against misinformation that has received less focus is logic-based inoculation (Banas & Miller, 2013): preemptive refutation of the critical thinking errors that make us vulnerable to misinformation in the first place. For example, Biddlestone and colleagues (2022) have used prebunking to reduce susceptibility to the cognitive bias of conjunction errors, subsequently reducing conspiracy beliefs and improving truth discernment between real and fake news headlines. This approach includes a potential added benefit of broad applicability to many unrelated pieces of misinformation due to its relevance to general critical thinking, potentially ensuring the broader scalability of prebunking messages (see also Biddlestone et al., 2025).

## Prebunking misinformation at scale

One way that researchers have administered these messages is through serious games that promote active inoculation: placing the player in the position of someone spreading disinformation online, being rewarded with points for using weakened doses of specific manipulation tactics (e.g., Basol et al., 2020; Neylan et al., 2023). These active inoculation interventions may be particularly effective due to their immersive nature (Basol et al., 2021), conducive to the internalization of the manipulation tactics in question (see also Green et al., 2022). However, there may be a trade-off between the improved internalization afforded by active inoculation and the ease of administration that passive inoculations enjoy. That is, to play inoculation games, participants are required to be motivated to play the games through to the end and to understand their content.

The development and dissemination of educational videos helps circumvent issues associated with the motivated administration of inoculation interventions by increasing the reach to viewers. This can be done through promotional campaigns that present the videos as adverts or on social media feeds, arguably the most dominant mode of media consumption on social media platforms. For example, short prebunking videos can be presented in the (non-skippable) ad space on YouTube and scaled to millions of people online before they see potential misinformation (Jigsaw, 2023; Roozenbeek et al., 2022).

## The constant need for new video content

As discussed, many options exist for campaigns looking to improve the reach of prebunking videos in specific contexts. However, some scholars argue that due to the constantly changing nature of misinformation online, misinformation interventions should be treated as a lifelong education tool requiring constant reinforcement alongside societal developments (e.g., Jaeger & Taylor, 2021). Furthermore, there are still many manipulation tactics that intervention

videos are yet to prebunk. For example, despite evidence suggesting that conspiracy theories thrive in times of societal crisis—such as the COVID-19 pandemic and climate crisis (e.g., Biddlestone, Azevedo, & van der Linden, 2022; Biddlestone et al., 2025; van Prooijen & Douglas, 2017)—there is no available video content that can be easily distributed to prebunk conspiracy theorizing. The current work also provides the opportunity to investigate the possibility of “cross-protection” between videos—a phenomenon in which prebunking against one manipulation tactic increases the detection of other, unrelated tactics (e.g., Lewandowsky & Yesilada, 2021)—which has not been possible in work investigating the efficacy of other prebunking videos (e.g., Roozenbeek et al., 2022).

Here, we aim to extend the work laid out by Roozenbeek et al. (2022) across two studies, by developing and testing a new collection of inoculation videos. In the first study, we hope to demonstrate the efficacy of videos prebunking the manipulation tactics of polarization (divisive content), conspiracy theories (unverified or unverifiable claims of secret nefarious plots that challenge official explanations of events; see Douglas & Sutton, 2023), and fake experts (the use of false or exaggerated credentials to create a false impression of competence). We categorize this content as manipulation techniques due to their usage to distort individuals' beliefs on a topic without providing clear evidence or logical arguments. For example, even though real conspiracies do occur (e.g., the *Watergate Scandal*), the use of conspiracy theories to encourage a general pattern of paranoid and distrustful thinking is not evidence-based regardless of whether questioning the official explanations for events is warranted. Boudry (2022) argues for a priori skepticism insofar conspiracy theories are the epistemic equivalent of black holes in their extreme resilience to refuting evidence, meaning that it is ultimately impossible to tell whether a conspiracy theory was responsible for the uncovering of an actual conspiracy that turns out to be true (though more recent research has argued for distinguishing between implausible and more plausible conspiracy theories, see O'Mahony et al., 2024). Polarization can be seen in a similar way, a manipulation tactic that is often used by misinformers (e.g., through purposefully or accidentally presenting misleading data visualizations about group differences; see Hanel et al., 2019; Tartaglione & De-Wit, 2024) to produce misleading headlines that, among other things, depict rare instances as representative of the views or behaviors of an entire group.

In the second study, we hope to demonstrate the efficacy of logic-based inoculation videos prebunking the logical fallacies of whataboutism (deflection from one's own immorality by accusations of similar immorality among others), the straw man fallacy (refuting an exaggerated and/or false version of your opponent's argument), and the moving the goalposts fallacy (alteration of the type of evidence required to refute or support one's argument after it has been challenged). We categorize this content as logical fallacies due to its usage in arguments to obscure logical reasoning and exploit cognitive heuristics (see also Hruschka & Appel, 2023).

## The present research

In the current article, we present two experiments aiming to demonstrate the efficacy of three technique-based (Study 1) and three logic-based (Study 2) inoculation videos at improving discernment of relevant manipulative content from non-manipulative content presented in social media posts. In both studies, participants were randomly allocated to one of the three respective experimental conditions in which they viewed one of the inoculation videos, or the control group in which they watched an educational video about freezer burn. Next, following Roozenbeek et al. (2022), participants were presented with a combination of manipulative and non-manipulative social media posts and asked to rate (1) how manipulative they perceived each post to be, (2) how confident they were in their manipulativeness assessments, and (3) how likely they were to share each post with people in their network.

We hypothesized that participants in the respective experimental conditions would demonstrate significantly stronger manipulation discernment (i.e., rating manipulative posts as more manipulative relative to non-manipulative posts; see H1a, H2a, and H3a in both pre-registration documents), greater confidence in these manipulateness assessments (see H1b, H2b, and H3b in both pre-registration documents), and increased quality of their sharing decisions (i.e., reduced intentions to share manipulative posts relative to non-manipulative posts; see exploratory analyses sections in both pre-registration documents) than participants in the control conditions.

Alongside our main experimental measures, we also included theoretical measures of participants' retrospective reporting of how much they counterargued against the points made in the videos, the motivation to resist manipulation in the future, as well as variables capturing how often, how many times, and with how many people participants discussed manipulative content with 2 weeks prior to completing the study. These were intended to give us further insights into the psychological mechanisms of the interventions administered (e.g., whether inoculation through video content increases motivated defense against the threat of manipulation as it does in some other inoculation mediums; see Compton & Ivanov, 2012; Maertens, Roozenbeek, et al., 2025; van der Linden, 2022; see exploratory analyses sections in both pre-registration documents). Finally, alongside socio-demographics (e.g., education, political ideology), we also included additional measures of general misinformation susceptibility and prior conspiracy beliefs in both studies to explore whether the efficacy of the interventions was conditional on prior susceptibility to general or conspiracist misinformation.

## STUDY 1

In Study 1, we aimed to test the efficacy of three videos using technique-based inoculation to prebunk the manipulation tactics of polarization, conspiracy theories, and fake experts in improving manipulation discernment of relevant manipulative versus non-manipulative social media posts. We hypothesized that viewing one of the three videos would increase manipulation discernment of the respective manipulation tactic included in the video compared to the control group. However, we also intended to explore the possibility of cross-protection: increases in manipulation discernment of unrelated manipulation tactics.

Furthermore, we also hypothesized that participants in the experimental conditions would increase their social media sharing decisions for the relevant content compared to participants in the control group, and that they would also display higher confidence in their manipulateness assessments. We also included measures of reported counterarguing against the points made in the videos to test resistance against the videos' messages, motivated resistance against manipulation, how often participants discussed manipulative content recently, prior conspiracy beliefs, general misinformation susceptibility, and socio-demographics to explore whether the efficacy of the interventions was conditional on these factors. The sample size, measures, hypotheses, and analyses were fully pre-registered and can be found at [https://aspredicted.org/Y49\\_KQD](https://aspredicted.org/Y49_KQD).<sup>1</sup> Please note that while we pre-registered our analyses of manipulation discernment and confidence in two respective single models, we deviated from these pre-registered analyses in both studies on the suggestion of peer-reviewers to conduct separate analyses for each manipulation tactic. We then included the effects of all videos on overall manipulation discernment, accounting for item

<sup>1</sup>Please note that after submitting the initial pre-registration but before collecting data, we decided to change our intended sample source from US to UK participants due to the British accent of the narrator in the videos. The original identical pre-registration that specified US participants as the intended sample can be found here [https://aspredicted.org/WJT\\_VQV](https://aspredicted.org/WJT_VQV).



categories, in a series of linear mixed effects models to explore whether findings could be accounted for by individual item or item category variance (see Supplement for full details of these analyses; Sections 7 and 17). All video and survey stimuli can be viewed in the on-line repository at <https://osf.io/8g7jcl>.

## Methods

### Participants

A total of 1,612 responses from UK *Prolific Academic* workers were collected and paid the minimum “fair” amount for their time. Once participants were removed for failing one of the attention checks (in line with our pre-registered exclusion criteria), the final sample size was 1,583, 772 women, 778 men,  $M_{\text{age}} = 39.93$ . There were 400 participants in the control condition, 401 participants in the polarization condition, 384 participants in the conspiracy theories condition, and 393 participants in the fake experts condition. Sensitivity power analysis using G\*Power indicated that with an assumed total sample size for four conditions ( $N = 1536$ ) all consisting of the number of participants in the condition with the lowest power (conspiracy theories condition  $N = 384$ ), a power of .80 was achieved to detect a small omnibus experimental effect between conditions,  $f = .08$ , and a small experimental effect between each experimental condition and the control group,  $d = .18$ .

### Design and procedure

After providing their consent, participants were randomly allocated to one of the four conditions. The scripts for each video were written by the authors and animated by *Studio You* and *Lens Change*. Each video was around 90 seconds long, with English subtitles embedded. Screenshots of each video can be seen in Figure 1, and the videos themselves can be viewed in the online repository: <https://osf.io/8g7jcl>.

In the conspiracy theories video, conspiracy theories are first defined as “a belief that some covert but influential entity is pulling the strings, and responsible for an unexplained event.” Next, an affective forewarning is shown with an exclamation mark graphic “Watch out, you are being targeted right now!” A microdose scenario is then presented in which the viewer is told they are only watching the video because “they” have traced your online behavior and are manipulating you into watching it. The preemptive refutation is then presented, explaining how conspiracy theories create an illusion of a secret plot of shady people who, in this example, are watching you. Additional information is provided about how people become vulnerable to (particularly false) conspiracy theories in times of fear and uncertainty. Finally, the viewer is reminded to keep a lookout for conspiracy theories online as another affective warning (usually the affective forewarning).



**FIGURE 1** Screenshots from the conspiracy theories (left), fake experts (centre), and polarization (right) videos.

Conspiracy theories


Manipulative




The bitcoin exchange rate is being manipulated by a small group of rich bankers. [#InvestigateNow](#)



BREAKING: Insurance companies are using your phone to track your fast food consumption.



Scientists discovered solution to greenhouse effect years ago but aren't allowed to publish it, report claims.



If y'all want to be blind to what's going on in society then go ahead. Don't call people crazy for seeking the actual truth instead of believing everything the government tells them.

Non-manipulative



British government considering digital-pound cryptocurrency.



Big tobacco always knew that cigarettes cause cancer, but they covered it up for profit.



Scientists claim solar and wind power are viable, cleaner energy alternatives to fossil fuels.



Oil company Exxon discovered the damage that fossil fuels would have on the planet in the 1950s, but funded media narratives to distract from this.

Fake experts


Manipulative




Famous economist exposes the pseudoscience behind central banks.



NASA scientists: The moon is actually just a reflection of the earth in the night sky.





Nutritionist claims that alcohol can actually improve liver function.





Historian suggests that theory of humans arriving as aliens from another planet is more likely than theory of evolution.


Non-manipulative





If the government prints money, this would contribute to increases in interest rates and inflation.



Contrary to popular opinion, it's actually not possible to view the Great Wall of China from space.



When you look at number of deaths and addiction rates, alcohol is more dangerous than cannabis.



Considering how unlikely it is for 1) species to develop, and 2) for them to be intelligent...the chances of aliens existing are slim.

FIGURE 2 Manipulative items (left) and their corresponding non-manipulative counterpart items (right) for each of the three manipulation tactics in Study 1.



Polarization

Manipulative

Non-manipulative

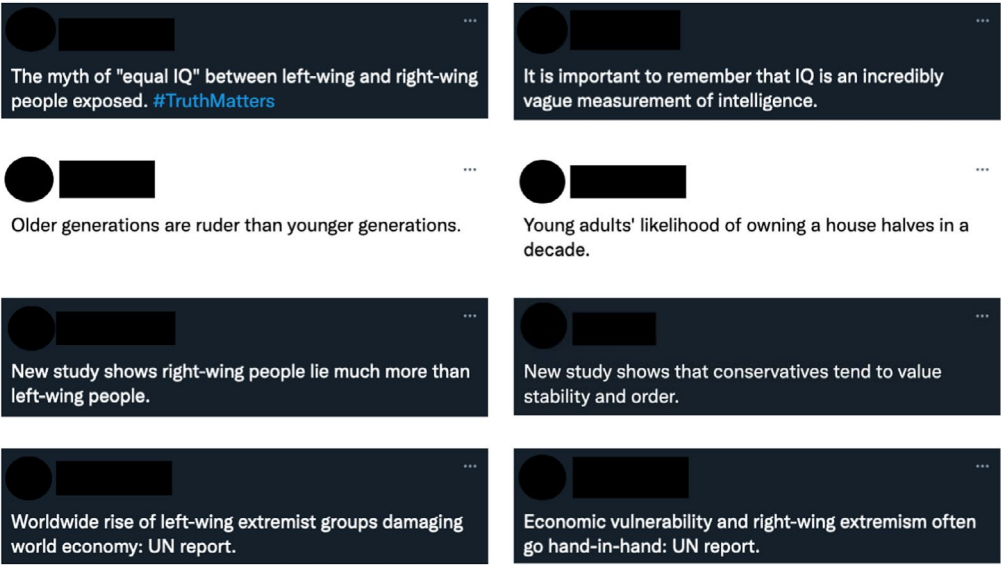


FIGURE 2 (Continued)

In the fake experts video, fake experts are first defined as “when people misuse or tout unverified credentials to achieve their agenda.” A microdose scenario is then presented in which Dr. Trusmi is described as having a “GhD in cashology” from the “University of Multilevel Marketing,” peddling business and health advice with other online gurus. The preemptive refutation is then presented, explaining how “expert” isn’t a protected term, allowing people to misuse or fake their credentials. Additional microdoses that are relevant to the real world are then provided, explaining that people engage in this tactic to make money, get invited onto TV shows and podcasts, and to achieve political goals. Finally, the viewer is reminded not to base their judgments on a single expert’s credentials, and instead to check the quality of the evidence and credentials of each expert as an affective warning (usually the affective forewarning).

In the polarization video, polarization is first defined as “when people deliberately seek to divide viewers into opposing groups.” A microdose scenario is then presented in which a pineapple pizza lover is reported to be caught in horrific, immoral, and possibly illegal acts with a pineapple pizza (see Figure 1). This story is used to push the narrative that pineapple pizza lovers are dangerous to, for example, children. The preemptive refutation is then presented, explaining how media manipulators could scour the web for rare instances of pineapple pizza lovers engaging in questionable behavior, using it to polarize people against pineapple pizza lovers by claiming this behavior is representative of the whole group. The narrator explains that disparaging language toward other groups is one of the strongest drivers of engagement on social media. Additional microdoses that are relevant to the real world are then provided, explaining that people engage in this tactic to generate social media engagement, increase polarization, or to anger an audience. Finally, the viewer is reminded that manipulative polarization may be occurring if content makes you feel upset and outraged at a particular group as an affective warning (usually the affective forewarning).

After watching the video, participants responded to an attention check question relating to what they saw in the video before proceeding to the item rating task. For the item rating

task, we created eight fake social media posts for each manipulation tactic, totaling 24 posts (see Figure 2). For each tactic, four of the posts contained manipulation using the respective tactic, and the four other posts were non-manipulative counterpart posts. Participants were randomly presented with 12 of the 24 total posts, seeing either the manipulative or non-manipulative counterpart of each post in randomized order, but never both counterparts. For the manipulative posts, these were intended to capture clear instances of the use of (a) conspiracy theories, (b) fake experts, and (c) polarization to make (a) unverified or unverifiable claims, (b) obscure a lack of clear evidence for claims, or (c) portray groups in a negative manner (e.g., morally) respectively. For the non-manipulative posts, these were intended to cover the same informational topic as their respective manipulative counterpart posts, without the use of a manipulation tactic to obscure or portray the information in a certain way.

To note, while some information in the manipulative posts may have held a kernel of truth (e.g., the manipulative polarization post about unequal IQs between left- and right-wingers is backed by some preliminary evidence; see Edwards et al., 2024), it is the raising of issues as a way to encourage judgmental perceptions between groups to amplify polarization that is intended to capture manipulation. For each post, participants were asked to rate the manipulateness of each post, their confidence in their manipulateness assessment of each post, and their intentions to share each post within their social network.

After completing the item rating task, participants were presented with the questions measuring their retrospective reporting of counterarguing against the points made in the video, motivation to resist manipulative content on social media in the future, as well as how often, how many times, and with how many people they discussed manipulative content with 2 weeks prior to completing the study in randomized order. Finally, participants reported their socio-demographics, general misinformation susceptibility (measured using the eight-item misinformation susceptibility test; MIST-8; Maertens, Götz, et al., 2023), and prior conspiracy beliefs (measured with the conspiracy mentality questionnaire; CMQ; Bruder et al., 2013), before being given the opportunity to report any feedback they may have wanted to offer and being fully debriefed.

## Ethics

Full ethical approval was obtained from the University of Cambridge Psychology Research Ethics Committee. PRE.2023.028.

## Measures

Response scales were measured from 1 (*strongly disagree*) to 7 (*strongly agree*) unless otherwise mentioned.

**Manipulation discernment** ( $M=1.41$ ,  $SD=1.29$ ) was measured by presenting participants with the single item statement “This post is manipulative” below each social media post. In line with previous research (e.g., Roozenbeek et al., 2022), manipulation discernment scores were calculated by subtracting the mean perceived manipulateness of the non-manipulative posts from the mean perceived manipulateness of the manipulative posts for each manipulation tactic so that higher scores indicated stronger manipulation discernment.

**Confidence in manipulateness assessments** was measured by presenting participants with the single item statement “I am confident in my assessment of the manipulateness of this post” below each of the manipulative ( $M=5.38$ ,  $SD=1.29$ ) and non-manipulative social media posts ( $M=5.17$ ,  $SD=.99$ ).

**Sharing discernment** ( $M=.70$ ,  $SD=1.05$ ) was measured by presenting participants with the single item statement “I would share this post with people in my network” below each of the manipulative ( $M=1.95$ ,  $SD=1.14$ ) and non-manipulative ( $M=2.65$ ,  $SD=1.36$ ) social media posts. To calculate a sharing discernment score, mean intentions to share the manipulative posts was subtracted from mean intentions to share the non-manipulative posts for each manipulation tactic so that higher scores indicated stronger sharing discernment.

**Counterarguing against the points made in the videos** ( $M=2.78$ ,  $SD=1.45$ ) was measured by presenting participants with the single item developed by Miller and colleagues (2013), asking them to indicate their thoughts while they were watching the video using a response scale from 1 (*I accepted all of the points made in the message*) to 7 (*I argued against all of the points made in the message*).

**Motivation to resist misinformation** ( $M=4.81$ ,  $SD=1.20$ ) was measured using an adapted three-item scale originally developed by Banas and Richards (2017) with the pre-amble “Thinking about the idea of [polarization/conspiracy theories/fake experts/manipulative posts] on social media...” and three items: “...motivates me to resist information,” “...I feel ready to argue against [polarising headlines/conspiracy theory headlines/fake expert headlines/manipulative headlines],” and “...makes me want to defend my attitudes against deceptive news.”

**How often participants discussed manipulative content** ( $M=2.18$ ,  $SD=1.46$ ) 2 weeks prior to completing the study was measured by asking participants “In the past two weeks, how often did you...” with the two questions “...talk about the issue of deceptive videos using [polarization/conspiracy theories/fake experts/manipulative content] with other people?” and “...talk about the issue of [polarization/conspiracy theories/fake experts/manipulative content] on social media?” using a response scale from 1 (never) to 7 (very often).

**The number of times participants discussed manipulative content** ( $M=1.10$ ,  $SD=1.46$ ) 2 weeks prior to completing the study was measured with a single item: “In the past two weeks, how many times did you talk about or discuss the issue of manipulative news (e.g., using [polarization/conspiracy theories/fake experts/emotional language])?” using a response scale from 0 to “More than 5.”

**The number of people participants discussed manipulative content with** ( $M=.96$ ,  $SD=1.28$ ) 2 weeks prior to completing the study was measured with the single item: “In the past two weeks, with how many people did you talk or discuss about the issue of manipulative news (e.g., using [polarization/conspiracy theories/fake experts/emotional language])?” using a response scale from 0 to “More than 5.”

**Prior conspiracy beliefs** ( $M=6.65$ ,  $SD=1.97$ ;  $\alpha=.87$ ) were measured with the five-item CMQ (Bruder et al., 2013; e.g., “[I think that...] ...politicians usually do not tell us the true motives for their decisions”) using a response scale from 1 (certainly not) to 11 (certain).

**General misinformation susceptibility** was measured with the MIST-8 (Maertens, Götz, et al., 2023), which presents four real news headlines (e.g., “Attitudes Toward EU Are Largely Positive, Both Within Europe and Outside It”;  $M=3.11$ ,  $SD=1.02$ ) and four fake news headlines (e.g., “Certain Vaccines Are Loaded with Dangerous Chemicals and Toxins”;  $M=3.20$ ,  $SD=.97$ ), and participants are required to indicate using a binary scale whether they believe each headline is *Real* or *Fake*. General truth discernment was calculated by summing the number of correctly detected real and fake news headlines ( $M=6.30$ ,  $SD=1.54$ ).

**Socio-demographics** were measured by asking participants to indicate their gender, age, highest level of completed education ( $M=4.86$ ,  $SD=1.44$ ) from 1 (No formal education above age 16) to 6 (Doctorate), their news consumption with one item asking “How often do you check the news?” ( $M=3.80$ ,  $SD=.92$ ) and their social media usage with an item asking “How often do you use social media (Facebook, YouTube, Twitter, Reddit, Instagram, etc.)?” ( $M=3.95$ ,  $SD=1.00$ ) both using a response scale from 1 (never) to 5 (all the time), and their self-placed political ideology ( $M=5.87$ ,  $SD=3.68$ ) using a response scale from 1 (very liberal) to 7 (very conservative).

## Results

In all one-way ANOVAs reported here, the four experimental conditions were added as the independent variable, and the dependent variable was altered depending on the analysis. If a significant main effect was indicated, post-hoc analysis with a Tukey correction for multiple comparisons was conducted to investigate the significant differences between conditions. The main findings are reported here, whereas the details of the results for counterarguing against the points made in the videos, motivation to resist manipulation, confidence in manipulateness assessments, sharing decisions, covariate analyses, moderation analyses, and details of the exploratory multilevel model analysis of item variation can be found in the Supplement (Sections 3–7). We ran robustness check analyses on the data file without attention check failures excluded (see Supplement, Section 1). These findings would only be reported here when they deviate from the main analyses, but all were near identical to the main analyses.

### Manipulation discernment—H1a, H2a, and H3a

**Polarization posts.** There were significant differences in manipulation discernment of the polarization posts between conditions,  $F(3, 1392)=6.43$ ,  $p < .001$ , such that manipulation discernment was significantly stronger in the polarization condition compared to the control,  $p = .043$ ,  $d = .20$ , 95% CI [.05, .35], and fake experts conditions,  $p < .001$ ,  $d = .32$ , 95% CI [.18, .47] (see Figure 3). These findings support H2a.

**Conspiracy theory posts.** There were also significant differences in manipulation discernment of the conspiracy theory posts between conditions,  $F(3, 1366)=3.76$ ,  $p = .011$ , such that manipulation discernment was significantly stronger in the conspiracy theories condition compared to the fake experts condition,  $p = .006$ ,  $d = .24$ , 95% CI [.09, .39], but not the control condition,  $p = .132$ ,  $d = .17$ , 95% CI [.02, .32] (see Figure 3). These findings do not support H1a.

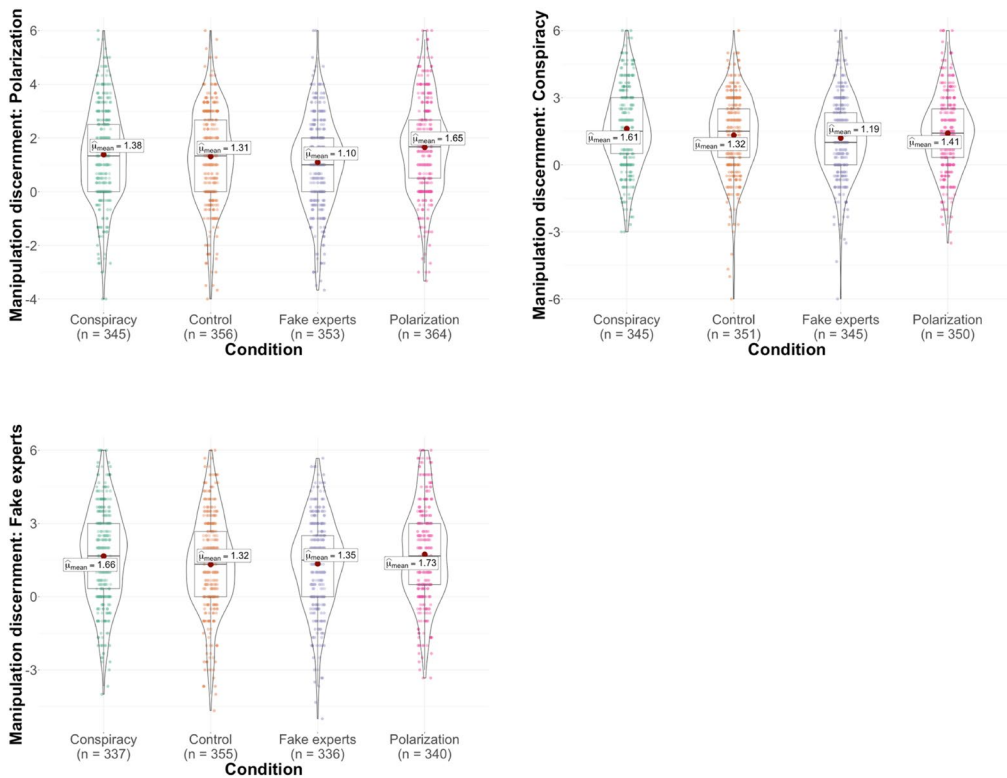
**Fake experts posts.** There were also significant differences in manipulation discernment of the fake experts posts between conditions,  $F(3, 1343)=3.87$ ,  $p = .009$ , such that manipulation discernment was significantly stronger in the polarization condition than in the fake experts,  $p = .045$ ,  $d = .20$ , 95% CI [.05, .35], and control conditions,  $p = .031$ ,  $d = .22$ , 95% CI [.07, .37], but there was no significant difference between the fake experts and control conditions,  $p = .999$ ,  $d = .02$ , 95% CI [−.13, .17] (see Figure 3). These findings do not support H3a.

### Manipulativeness assessments

**Non-manipulative polarization posts.** There were significant differences in manipulateness assessments of the non-manipulative polarization posts between conditions,  $F(3, 1464)=5.10$ ,  $p = .002$ , such that manipulateness assessments were significantly higher in the fake experts condition compared to the control condition,  $p = .001$ ,  $d = .28$ , 95% CI [.13, .42] (see Figure 4).

**Manipulative polarization posts.** There were also significant differences in manipulateness assessments of the manipulative polarization posts between conditions,  $F(3, 1479)=11.22$ ,  $p < .001$ , such that manipulateness assessments were significantly higher in the polarization condition compared to the control,  $p < .001$ ,  $d = .41$ , 95% CI [.27, .55], conspiracy theories,  $p = .001$ ,  $d = .28$ , 95% CI [.13, .42], and fake experts conditions,  $p < .001$ ,  $d = .30$ , 95% CI [.16, .44] (see Figure 4).

**Non-manipulative conspiracy theory posts.** There were significant differences in manipulateness assessments of the non-manipulative conspiracy theory posts between conditions,  $F(3, 1466)=3.83$ ,  $p = .010$ , such that manipulateness assessments were significantly higher in the



**FIGURE 3** Violin plots of the experimental effects of each condition on manipulation discernment of the polarization (top left), conspiracy theory (top right), and fake expert posts (bottom left).

fake experts,  $p = .015$ ,  $d = .22$ , 95% CI [.08, .37], and polarization conditions,  $p = .029$ ,  $d = .21$ , 95% CI [.07, .36], compared to the control condition (see Figure 5).

**Manipulative conspiracy theory posts.** There were also significant differences in manipulateness assessments of the manipulative conspiracy theory posts between conditions,  $F(3, 1451) = 6.42$ ,  $p < .001$ , such that manipulateness assessments were significantly higher in both the conspiracy theories,  $p < .001$ ,  $d = .30$ , 95% CI [.15, .44], and polarization conditions,  $p = .002$ ,  $d = .27$ , 95% CI [.13, .41], compared to the control condition (see Figure 5).

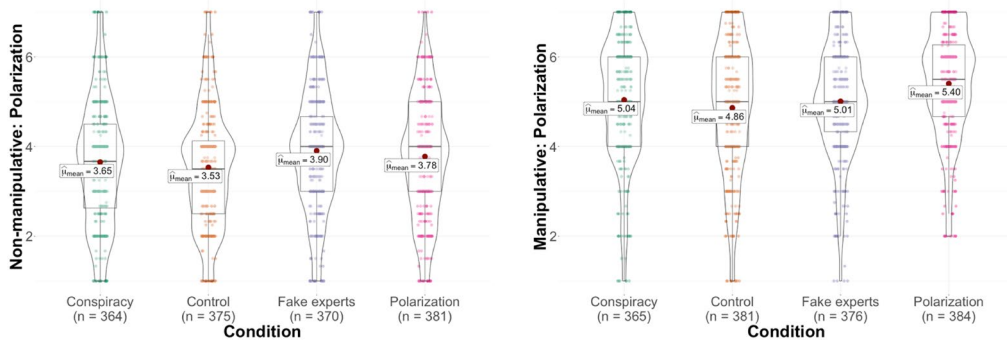
**Non-manipulative fake expert posts.** There were no significant differences in manipulateness assessments of the non-manipulative fake expert posts between conditions,  $F(3, 1440) = 2.17$ ,  $p = .090$  (see Figure 6).

**Manipulative fake expert posts.** There were significant differences in manipulateness assessments of the manipulative fake expert posts between conditions,  $F(3, 1454) = 5.99$ ,  $p < .001$ , such that manipulateness assessments were significantly higher in the fake experts,  $p = .035$ ,  $d = .21$ , 95% CI [.07, .35], polarization,  $p < .001$ ,  $d = .30$ , 95% CI [.16, .45], and conspiracy theories conditions,  $p = .005$ ,  $d = .26$ , 95% CI [.12, .40], compared to the control condition (see Figure 6).

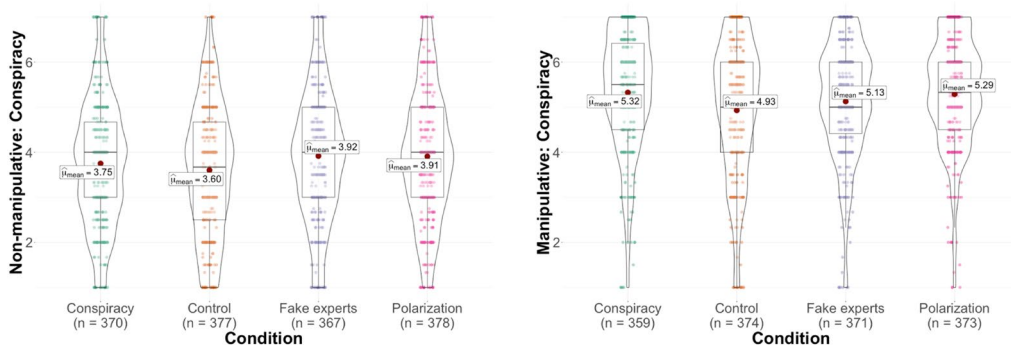
## Exploratory linear mixed effects models to test item variation

To explore variance based on individual items and item categories, a series of linear mixed effects models were analyzed (see Supplement, Section 7). Scores for the non-manipulative





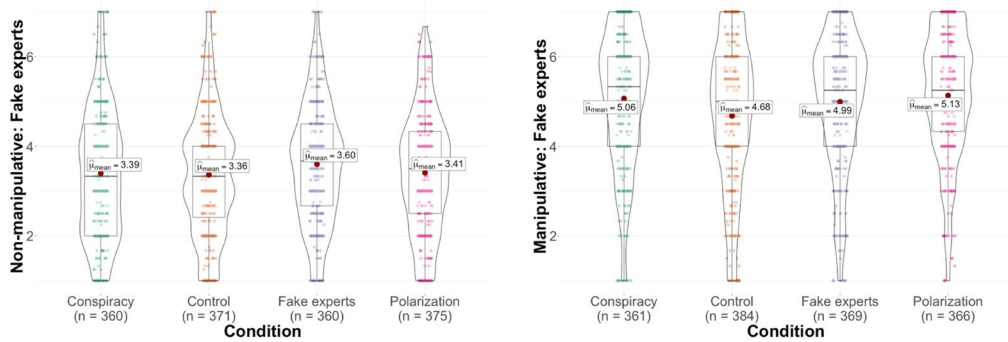
**FIGURE 4** Violin plots of the experimental effects of each condition on manipulativeness perceptions of the non-manipulative (left) and manipulative (right) polarization posts.



**FIGURE 5** Violin plots of the experimental effects of each condition on manipulativeness perceptions of the non-manipulative (left) and manipulative (right) conspiracy theory posts.

posts were reverse coded so that higher scores on all items indicated manipulation discernment. These models indicated that the conspiracy theories,  $b = .13$ ,  $SE = .06$ ,  $t(4,595) = 1.97$ ,  $p = .049$ , polarization,  $b = .20$ ,  $SE = .06$ ,  $t(4,852) = 3.17$ ,  $p = .002$ , and fake experts videos (vs. control),  $b = .13$ ,  $SE = .06$ ,  $t(4,808) = 2.01$ ,  $p = .045$ , each significantly improved manipulation discernment of their respective manipulation tactics when analyzed separately, and that individual items and item categories accounted for a negligible amount of the variance in these findings (see Supplement, Section 7). Considerable residual variance was also indicated in each model, suggesting that the majority of variance was explained by factors that were unaccounted for in the models (see Supplement, Section 7). In a model containing all three videos (vs. control) as simultaneous predictors of overall manipulation discernment, only the conspiracy theories,  $b = .12$ ,  $SE = .04$ ,  $t(1,603) = 2.90$ ,  $p = .004$ , and polarization videos (vs. control),  $b = .12$ ,  $SE = .04$ ,  $t(1,607) = 2.86$ ,  $p = .004$ , significantly improved overall manipulation discernment, whereas the effect for the fake experts video (vs. control) was non-significant,  $b = .06$ ,  $SE = .04$ ,  $t(1,572) = 1.34$ ,  $p = .182$  (see Supplement, Section 7). Once again, individual items and item categories accounted for a negligible amount of variance, variance = .02,  $SD = .16$ , which was instead explained by factors unaccounted for in the model, residual variance = 4.61,  $SD = 2.15$ .





**FIGURE 6** Violin plots of the experimental effects of each condition on manipulativenness perceptions of the non-manipulative (left) and manipulative (right) fake expert posts.

## Discussion

In Study 1, we set out to determine whether our inoculation videos developed to prebunk the manipulation tactics of polarization, conspiracy theories, and fake experts significantly increased manipulation discernment of social media posts that used these tactics from ones that did not, compared to the control group. The polarization inoculation video significantly increased manipulation discernment of the manipulative polarization posts from the non-manipulative ones, compared to the control group (supporting H2a),  $d = .20$ , and this appeared to be due to an increase in the detection of the manipulative polarization posts,  $d = .40$ , with no significant difference in the detection of non-manipulative polarization posts compared to the control group. Despite no increase in confidence in these manipulativenness assessments or sharing decisions after viewing the polarization inoculation video compared to the control group, the polarization inoculation video also increased manipulation discernment of the fake expert posts, and manipulativenness assessments of both manipulative conspiracy theory and fake expert posts compared to the control group. These findings suggest that the polarization inoculation video was largely effective at improving manipulation discernment of polarizing social media posts and demonstrated some cross-protection against manipulative conspiracy theory and fake expert posts.

While the conspiracy theories inoculation video did not significantly increase manipulation discernment of the manipulative versus non-manipulative conspiracy theory posts (refuting pre-registered hypothesis H1a), it did significantly increase the detection of manipulative conspiracy theory posts compared to the control group,  $d = .28$ . Furthermore, this was reflected in higher confidence in these manipulativenness assessments (supporting H1b; see Supplement, Section 4). Interestingly, there was evidence of cross-protection from the conspiracy theories condition in the higher manipulativenness assessments of manipulative fake expert posts and corresponding increases in confidence in these manipulativenness assessments compared to the control group. However, the conspiracy theories inoculation video did not significantly increase sharing decisions of the conspiracy theory posts compared to the control group. Despite notable floor effects for sharing intentions of all non-manipulative and manipulative posts, these findings suggest that in contrast to the polarization inoculation video, the conspiracy theories inoculation video was ineffective at improving relevant discernment, but still led to higher detection—and confidence in this detection—of conspiracy theory and fake expert manipulation.

Despite the lack of increased manipulation discernment between manipulative and non-manipulative fake expert posts after watching the fake experts inoculation video (refuting H3a), the fake experts video significantly increased manipulativenness assessments of manipulative fake expert posts compared to the control group,  $d = .20$ . However, there was no significant difference

in manipulateness assessments of the non-manipulative fake expert posts after watching the fake expert video compared to the control group, and confidence and sharing decisions were not significantly different compared to the control group. Once again, these findings suggest that the fake experts inoculation video was only effective at improving the detection of fake expert manipulation, but not the ability to discern this information from non-manipulative fake expert posts, or indeed either the decisions to share or confidence in detecting it.

Exploratory linear mixed effects models indicated that all three videos (vs. control) improved manipulation discernment of their respective manipulation tactics when accounting for the variance in individual items and item categories. However, when all three videos were entered into the same model, only the polarization and conspiracy theories videos (vs. control) improved overall manipulation discernment. Since variance accounted for by the items was negligible, it appears that the variation in findings is likely caused by factors that were not accounted for in these models. So far, it appears that the inoculation videos developed were largely effective at improving the detection of relevant manipulation. Furthermore, in some cases, there was evidence of cross-protection against manipulation in other domains. Therefore, we can conclude that the effect of the videos is likely to increase the detection of relevant manipulation. However, manipulateness assessments of the non-manipulative polarization and conspiracy theory posts were significantly *higher* in the fake experts condition compared to the control. Coupled with the lack of higher manipulation discernment of relevant content in the fake experts condition compared to the control group, this suggests that the fake experts condition could have increased unrelated skepticism as well as discernment.

## STUDY 2

In Study 2, we set out to run a similar experiment to Study 1 in a different context. Specifically, we aimed to test the efficacy of inoculation videos developed to prebunk the logical fallacies of whataboutism, the straw man fallacy, and the moving the goalposts fallacy. In this way, Study 1 investigated the effects of technique-based inoculation videos on manipulation discernment, whereas Study 2 looks at the effects of logic-based inoculation videos on fallacy discernment. The design and procedure were nearly identical to Study 1, replacing the videos and item rating task with content relating to the logic-based inoculation. The sample for Study 2 was also changed to US citizens instead of the United Kingdom. Furthermore, to address recent efforts to further understand the efficacy of inoculation interventions on different populations, we pre-screened participants so that roughly half were Republican and the other half Democrat voters. The sample size, measures, hypotheses, and analyses were fully pre-registered and can be found at [https://aspredicted.org/NDK\\_RFQ](https://aspredicted.org/NDK_RFQ). Please note that while we pre-registered our analyses of fallacy discernment and confidence in two respective single models, we deviated from these pre-registered analyses to conduct separate analyses for each fallacy. We then included the effects of all videos on overall fallacy discernment, accounting for item categories, in a series of linear mixed effects models to explore whether findings could be accounted for by individual item or item category variance (see Supplement for full details of these analyses, Section 17). All videos and survey stimuli can be viewed in the online repository at <https://osf.io/8g7jc/>.

## Methods

### Participants

A total of 1,607 responses were collected from US *Prolific Academic* workers, who were paid the minimum “fair” amount for their time. Once participants were removed for failing both

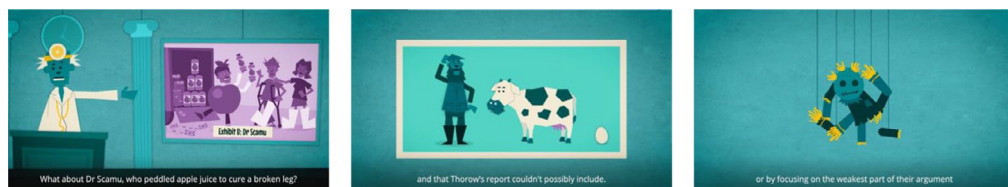
attention checks, the final sample was 1,603,  $N_{\text{Republicans}} = 800$ ,  $N_{\text{Democrats}} = 803$ , 784 women, 797 men,  $M_{\text{age}} = 43.51$ . There were 397 participants in the whataboutism condition, 403 participants in the straw man fallacy condition, 397 participants in the moving the goalposts fallacy condition, and 406 participants in the control condition. Sensitivity power analysis using G\*Power indicated that with an assumed total sample size for four conditions ( $N = 1,588$ ) all consisting of the number of participants in the condition with the lowest power (whataboutism and straw man fallacy conditions  $N = 397$  respectively), a power of .80 was achieved to detect a small omnibus effect between conditions,  $f = .08$ , and a small experimental effect between each experimental condition and the control,  $d = .18$ .

## Design and procedure

After providing their consent, participants were randomly allocated to one of the four experimental conditions. The scripts for the videos were written by the authors and animated by *Studio You* and *Lens Change*. Each video was around 1 minute long, with English subtitles embedded. Screenshots of each video can be seen in Figure 7, and the videos themselves can be viewed in the online repository: <https://osf.io/8g7jc/>.

In the whataboutism video, a microdose scenario is established in which Dr. Trusmi is on trial for medical malpractice. Dr. Trusmi is accused of attempting to market grapefruit juice as a cure for a broken arm. He pleads innocent, arguing that he has only been doing the same as Dr. Scamu, who has not been arrested despite peddling apple juice as a cure for a broken leg. In response, the judge states “Pointing a finger at another is no legal defence. GUILTY AS CHARGED!” For the preemptive refutation, the narrator then explains how Dr. Trusmi used whataboutism, explaining its usage in this context. Other, microdoses of whataboutism more relevant to real life are then provided: world leaders deflecting blame for human rights violations by pointing at another country's behavior and lying politicians deflecting attention to their rivals' lies. Finally, the viewer is reminded to keep a lookout for whataboutism online as an affective warning (usually the affective forewarning).

In the moving the goalposts fallacy video, a microdose scenario is established in which a contestant on a gameshow called *Facts Galore* argues that cows cannot lay eggs, one reason being they wouldn't be able to produce enough calcium to form an eggshell. Their opponent responds by claiming this argument is insufficient because it doesn't refer to a farmer's eyewitness account 2 weeks prior of a cow laying an egg. For the preemptive refutation, the narrator then mentions how the second contestant used the moving the goalposts fallacy, explaining how they referred to unverified information that could have been made up, and occurred too recently for the original contestant's argument to include. Other microdoses of the moving the goalposts fallacy relevant to politics and international relations are then provided: political figures sometimes dismiss evidence against them as they point to new information that could be unverified. Finally, the viewer is reminded to keep a lookout for the moving the goalposts fallacy online as an affective warning (usually the affective forewarning).



**FIGURE 7** Screenshots from the whataboutism (left), moving the goalposts fallacy (centre), and straw man fallacy (right) videos.

In the straw man fallacy video, a microdose scenario is established in which two politicians are debating over kids skateboarding in public areas. The first politician states that their policy would be to leave these kids alone unless they are committing a crime. In response, the other politician states “Did you hear that people? My opponent supports old ladies being run over by idiot teenagers on skateboards.” For the preemptive refutation, the narrator then mentions how the politician used the straw man fallacy, explaining how the first politician never said anything about old ladies being run over, but the other tried to make it look as if they did. Other microdoses of the straw man fallacy relevant to politics and international relations are then provided: manipulators completely misrepresenting the opponent's position, or by focusing on the weakest part of their argument to discredit it. Finally, the viewer is reminded to keep a lookout for the straw man fallacy online as an affective warning (usually the affective forewarning).

After watching the video, participants responded to an attention check question relating to what they saw in the video before proceeding to the item rating task. For the item rating task, we again created eight faux social media posts for each fallacy, totaling 24 posts (see [Figure 8](#)). For each fallacy, four of the posts contained manipulation using the respective fallacy, and the four other posts were non-fallacious counterpart posts (see [Figure 8](#)). Participants were first presented with four of the respective fallacious or non-fallacious counterpart posts in randomized order, and then eight of the fallacious or non-fallacious counterpart posts for the other two fallacies in randomized order. Participants were only ever presented with one of the two counterparts for each post, and for each post, they were asked to rate the manipulativeness, their confidence in these manipulativeness assessments, and their intentions to share each post within their social network.

After completing the item rating task, participants were presented with the questions measuring their motivations to resist misinformation on social media, alongside how often they had discussed manipulative content, how many times they had discussed manipulative content, and with how many people they discussed manipulative content in the 2 weeks prior to completing the study, as well as their retrospective reporting of how much they counterargued against the points made in the video in randomized order. Finally, participants reported their socio-demographics, general misinformation susceptibility (measured with the MIST-8; Maertens, Götz, et al., 2023), and prior conspiracy beliefs (measured with the CMQ; Bruder et al., 2013), before being given the opportunity to report any feedback they may have and being fully debriefed.

## Ethics

Full ethical approval was obtained from the University of Cambridge Psychology Research Ethics Committee. PRE.2023.028.

## Measures

All measures were identical to those used in Study 1 (see Supplement for details, [Section 8](#)).

## Results

The analytic strategy for Study 2 follows the same rationale as Study 1.

Whataboutism

Fallacious

Non-fallacious

Great, so we just focus on solving police brutality and just forget about all the officers killed in the line of duty...

Fatal police shootings in the US are increasing, but researchers disagree about the reasons why.

Okay so what's the idea, we increase funding for public schools and completely ignore privately funded schools?

Public schools are funded by local, state, or federal government while private schools are generally funded through tuition paid by the students.

Hilarious to me that the Iran gov. is being dragged through the mud for human rights abuses by THE US! Sure bud "cough" Guantanamo "cough" Abu Ghraib...

Guantanamo Bay and Abu Ghraib are sites where US military officials are alleged to have engaged in human rights abuses.

People obsessed with exploitation in the porn industry make me LOL. As if exploitation in sweatshops and the service industry isn't a problem...

Exploitation of women and men in the porn industry is common, according to a new report.

Moving the goalposts fallacy

Fallacious

Non-fallacious

I mean, maybe gun control reduces violent crime but it won't end gang violence, will it?

Personal protection is one of the most common reasons US adults give for owning firearms.

A cap on house prices would make houses more affordable for some, but there would still be homeless people.

A cap on house prices is being discussed as a potential solution for the rising cost of living.

I agree that the COVID-19 vaccines have reduced hospitalization rates, but they didn't prevent people from getting infected.

COVID-19 vaccines have reduced hospitalization rates by about 80-90%, according to a paper in The Lancet.

Supplying Ukraine with weapons will help them defend themselves against Russia I guess, but it won't fix decades of anti-Western attitudes in Russia, Iran and the Middle East

Iran has begun supplying Russia with military drones to be used in Ukraine.

FIGURE 8 Fallacious items (left) and their corresponding non-fallacious counterpart items (right) for each of the three logical fallacies in Study 2.

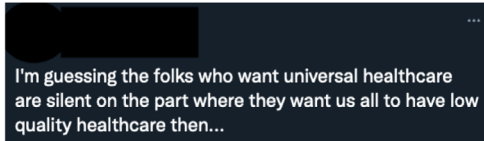


## Straw man fallacy

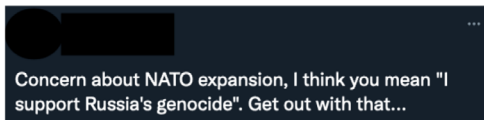
## Fallacious



Ah yea "invest in bike lanes", just say what you're really thinking and tell us you want cars banned!



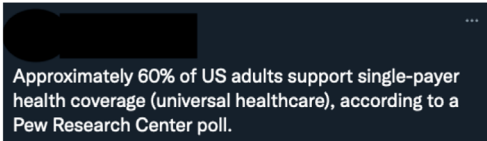
If you want voter ID laws you want to stop minorities from voting, full stop.



## Non-fallacious



Bike lanes are becoming more and more common throughout the US as an alternative to cars.



Voter identification is required in US states such as Georgia, Arkansas, Indiana, and Kansas



FIGURE 8 (Continued)

## Fallacy discernment—H1a, H2a, and H3a

**Whataboutism posts.** There were significant differences in fallacy discernment of the whataboutism posts between conditions,  $F(3, 1588) = 9.67, p < .001$ , such that fallacy discernment was significantly stronger in the moving the goalposts,  $p = .032, d = .19, 95\% \text{ CI } [.05, .33]$ , and straw man fallacy conditions compared to the control condition,  $p < .001, d = .37, 95\% \text{ CI } [.23, .51]$ . Furthermore, fallacy discernment of the whataboutism posts was significantly stronger in the straw man condition compared to the whataboutism,  $p = .011, d = .22, 95\% \text{ CI } [.08, .36]$ , and moving the goalposts conditions,  $p = .045, d = .19, 95\% \text{ CI } [.05, .32]$ . However, there was no significant difference in fallacy discernment between the whataboutism and control conditions,  $p = .121, d = .16, 95\% \text{ CI } [.02, .29]$  (see Figure 9). These findings do not support H3a.

**Moving the goalposts fallacy posts.** There were significant differences in fallacy discernment of the moving the goalposts fallacy posts between conditions,  $F(3, 1588) = 6.97, p < .001$ , such that fallacy discernment was significantly stronger in the straw man fallacy,  $p = .007, d = .23, 95\% \text{ CI } [.09, .37]$ , and whataboutism conditions,  $p = .001, d = .25, 95\% \text{ CI } [.11, .39]$ , compared to the control condition, but there was no significant difference in fallacy discernment between the moving the goalposts fallacy and control conditions,  $p = .962, d = .04, 95\% \text{ CI } [-.10, .18]$  (see Figure 9). These findings do not support H1a. Furthermore, fallacy discernment of the moving the goalposts fallacy posts was significantly lower in the moving the goalposts fallacy condition compared to the straw man fallacy,  $p = .034, d = .19, 95\% \text{ CI } [.05, .33]$ , and whataboutism conditions,  $p = .008, d = .21, 95\% \text{ CI } [.07, .35]$ .

**Straw man fallacy posts.** There were significant differences in fallacy discernment of the straw man fallacy posts between conditions,  $F(3, 1588) = 4.42, p = .004$ , such that fallacy



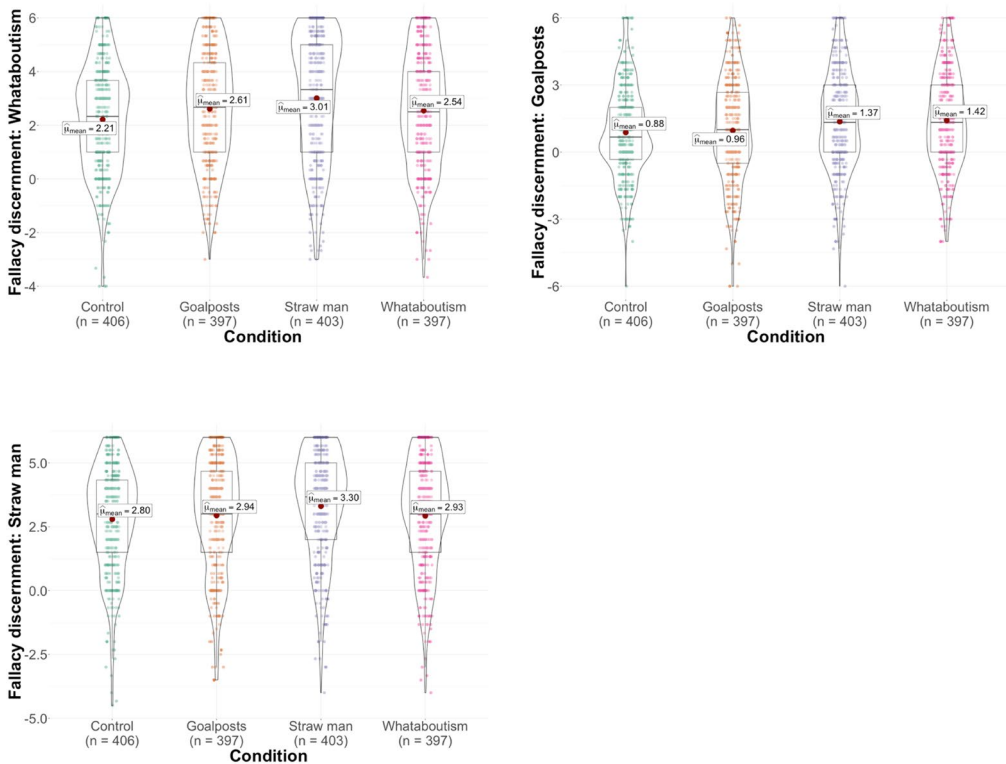
discernment was significantly stronger in the straw man fallacy condition compared to the control condition,  $p = .003$ ,  $d = .23$ , 95% CI [.09, .37] (see Figure 9). These findings support H2a.

## Manipulativeness assessments

**Non-fallacious whataboutism posts.** There were no significant differences in manipulateness assessments of the non-fallacious whataboutism posts between conditions,  $F(3, 1588) = 1.40$ ,  $p = .242$ .

**Fallacious whataboutism posts.** There were significant differences in manipulateness assessments of the fallacious whataboutism posts between conditions,  $F(3, 1588) = 15.73$ ,  $p < .001$ , such that manipulateness assessments were significantly higher in the whataboutism,  $p = .013$ ,  $d = .22$ , 95% CI [.08, .36], moving the goalposts,  $p < .001$ ,  $d = .35$ , 95% CI [.21, .48], and straw man fallacy conditions,  $p < .001$ ,  $d = .46$ , 95% CI [.32, .60], compared to the control condition (see Figure 10). Furthermore, manipulateness assessments of the fallacious whataboutism posts were significantly higher in the straw man fallacy condition compared to the whataboutism condition,  $p = .003$ ,  $d = .24$ , 95% CI [.10, .38].

**Non-fallacious moving the goalposts fallacy posts.** There were significant differences in manipulateness assessments of the non-fallacious moving the goalposts fallacy posts between conditions,  $F(3, 1588) = 7.01$ ,  $p < .001$ , such that manipulateness assessments were significantly higher in the moving the goalposts fallacy condition compared to the whataboutism,  $p < .001$ ,



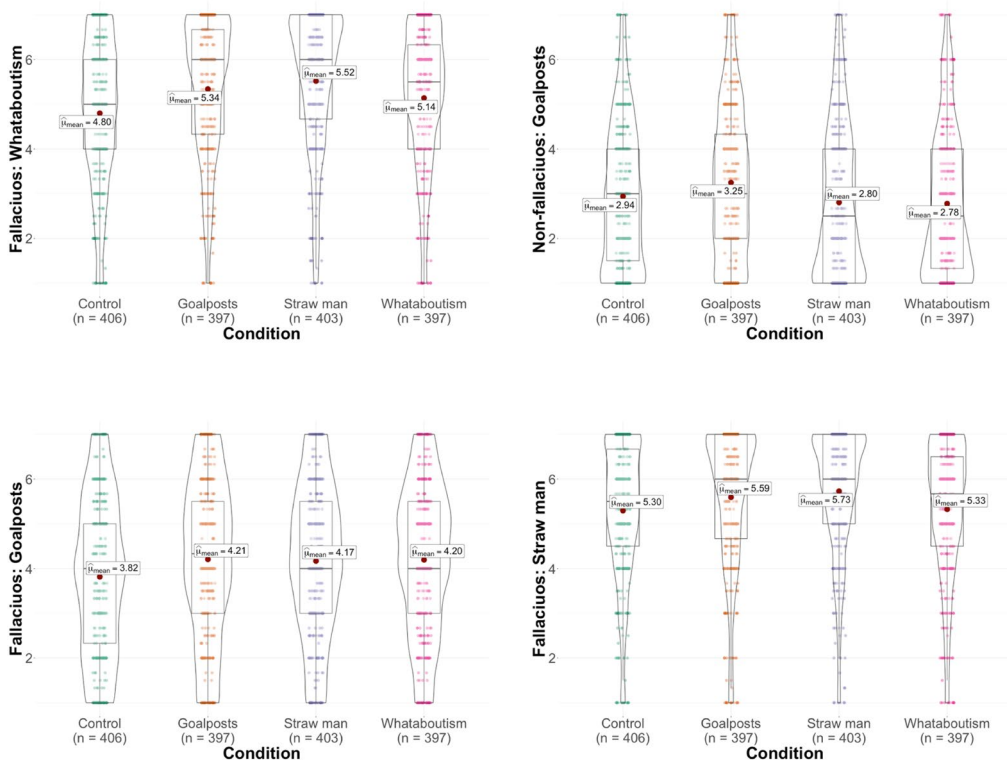
**FIGURE 9** Violin plots of the experimental effects of each condition on fallacy discernment of the whataboutism (top left), moving the goalposts fallacy (top right), and straw man fallacy posts (bottom left).

$d = .29$ , 95% CI [.15, .43], straw man fallacy,  $p < .001$ ,  $d = .27$ , 95% CI [.13, .41], and control conditions,  $p = .035$ ,  $d = .19$ , 95% CI [.05, .35] (see Figure 10).

**Fallacious moving the goalposts fallacy posts.** There were significant differences in manipulateness assessments of the fallacious moving the goalposts fallacy posts between conditions,  $F(3, 1588) = 4.98$ ,  $p = .002$ , such that manipulateness assessments were significantly higher in the moving the goalposts fallacy,  $p = .007$ ,  $d = .23$ , 95% CI [.09, .37], whataboutism,  $p = .006$ ,  $d = .22$ , 95% CI [.08, .36], and straw man fallacy conditions,  $p = .017$ ,  $d = .21$ , 95% CI [.07, .34], compared to the control condition (see Figure 10).

**Non-fallacious straw man fallacy posts.** There were significant differences in manipulateness assessments of the non-fallacious straw man fallacy posts between conditions,  $F(3, 1588) = 2.68$ ,  $p = .046$ . However, post-hoc analysis indicated that there were no significant differences in manipulateness assessments between conditions.

**Fallacious straw man fallacy posts.** There were significant differences in manipulateness assessments of the fallacious straw man fallacy posts between conditions,  $F(3, 1588) = 7.69$ ,  $p < .001$ , such that manipulateness assessments were significantly lower in the control condition compared to the straw man,  $p < .001$ ,  $d = .29$ , 95% CI [.15, .43], and moving the goalposts fallacy conditions,  $p = .028$ ,  $d = .19$ , 95% CI [.06, .34] (see Figure 10). Furthermore, manipulateness assessments of the fallacious straw man fallacy posts were significantly higher in the straw man fallacy condition compared to the whataboutism condition,  $p = .001$ ,  $d = .27$ , 95% CI [.13, .41].



**FIGURE 10** Violin Plots of the Experimental Effects of Each Condition on Manipulateness Assessments of the Fallacious Whataboutism posts (Top Left), Non-Fallacious Moving the Goalposts Fallacy Posts (Top Right), Fallacious Moving the Goalposts Fallacy Posts (Bottom Left), and Fallacious Straw Man Fallacy Posts (Bottom Right).

## Exploratory linear mixed effects models to test item variation

To explore variance based on individual items and item categories, a series of linear mixed effects models were analyzed (see Supplement, [Section 17](#)). These models indicated that the whataboutism,  $b = .11$ ,  $SE = .05$ ,  $t(801.01) = 2.06$ ,  $p = .040$ , and straw man videos (vs. control),  $b = .13$ ,  $SE = .05$ ,  $t(9,705) = 2.81$ ,  $p = .005$ , significantly improved fallacy discernment of their respective fallacies when analyzed separately, and that individual items and item categories only accounted for a moderate amount of the variance in these findings (see Supplement, [Section 17](#)). In contrast, the effect of the moving the goalposts condition (vs. control) did not significantly improve fallacy discernment of the moving the goalposts content,  $b = .06$ ,  $SE = .05$ ,  $t(801.03) = 1.02$ ,  $p = .306$ . Variance accounted for by the individual participant responses was moderate in this model,  $\text{variance} = .22$ ,  $SD = .47$ . Considerable residual variance was indicated in each model, suggesting that the majority of variance was explained by factors that were unaccounted for (see Supplement, [Section 17](#)). In a model containing all three videos (vs. control) as simultaneous predictors of overall fallacy discernment, only the straw man video (vs. control) significantly improved overall fallacy discernment,  $b = .14$ ,  $SE = .04$ ,  $t(1,599) = 3.19$ ,  $p = .001$ , whereas the effects for the whataboutism,  $b = .04$ ,  $SE = .04$ ,  $t(1,599) = 1.00$ ,  $p = .319$ , and moving the goalposts videos (vs. control),  $b = .04$ ,  $SE = .04$ ,  $t(1,599) = .87$ ,  $p = .386$ , were non-significant (see Supplement, [Section 17](#)). Individual items and item categories accounted for a notable amount of variance in this model,  $\text{variance} = 2.12$ ,  $SD = 1.46$ , which was also explained by factors unaccounted for in the model,  $\text{residual variance} = 2.65$ ,  $SD = 1.63$ .

## Discussion

In Study 2, we set out to determine whether our inoculation videos developed increased discernment of social media posts that contained the three logical fallacies from those that did not. Mixed findings were obtained across the videos. While the whataboutism inoculation video did not significantly increase fallacy discernment of the whataboutism posts compared to the control group (refuting H3a),  $d = .16$ , it did increase detection of the posts containing fallacious whataboutism content compared to the control group,  $d = .22$ , without altering perceptions of the non-fallacious whataboutism posts. Nevertheless, the whataboutism inoculation video did not appear to increase confidence in manipulativeness assessments (refuting H3b) or sharing decisions regarding the whataboutism posts (see Supplement, [Sections 14](#) and [15](#)).

With regard to the moving the goalposts fallacy inoculation video, fallacy discernment of the moving the goalposts fallacy posts was not increased relative to the control group (refuting H2a),  $d = .04$ , and manipulativeness assessments of the posts containing both fallacious,  $d = .22$ , and non-fallacious moving the goalposts fallacy content,  $d = .19$ , were significantly higher than in the control condition. This suggests that the moving the goalposts fallacy inoculation video increased overall distrust toward the relevant content rather than fallacy discernment. Furthermore, confidence in manipulativeness assessments (refuting H2b) and sharing decisions were not increased by the moving the goalposts fallacy inoculation video compared to the control group despite some higher detection of fallacious whataboutism and straw man posts (see Supplement, [Sections 14](#) and [15](#)).

The straw man fallacy inoculation video showed more promise, improving discernment of posts containing this fallacy from ones that did not compared to the control group (supporting H1a),  $d = .20$ . Furthermore, the straw man fallacy inoculation video increased confidence in manipulativeness assessments of posts containing the straw man fallacy (supporting H1b) and increased sharing discernment of these posts. Finally, the straw man fallacy inoculation video provided impressive cross-protection against the whataboutism and moving the Goalposts

fallacy posts, improving fallacy discernment and confidence in manipulateness assessments of their content.

Exploratory linear mixed effects models indicated that the whataboutism and straw man videos (vs. control) improved fallacy discernment of their respective fallacies when accounting for the variance in item categories. However, when all three videos were entered into the same model, only the straw man video (vs. control) improved fallacy discernment. Since variance accounted for by the item categories was negligible-to-moderate, it appears that the variation in findings is likely caused by factors that were not accounted for in these models.

Overall, these findings suggest that the straw man fallacy inoculation video is likely to confer higher abilities to detect the straw man fallacy and other fallacious content online. While the effects of the whataboutism inoculation video were less clear, it still increased the detection of fallacious whataboutism content without altering perceptions of non-fallacious content. Therefore, we can conclude that the message in this video is still somewhat educational despite providing less general cross-protection against logical fallacies than the straw man fallacy inoculation video. Finally, the moving the goalposts fallacy inoculation video appeared to increase distrust of the relevant content. As a result, this video is less promising with regard to the intended efficacy of these videos to specifically increase viewers' abilities to distinguish relevant fallacious content from non-fallacious content. That said, the lack of efficacy of this video is not without its strengths. Improving skepticism (i.e., distrust) has positive implications with regard to the motivation to seek out and detect when one is being manipulated. However, more work is needed to determine whether the effects of this video are simply less promising than the others or, worse, harmful to the cause of educating viewers. Participants reported much higher counterarguing against the points made in the moving the goalposts video compared to the other conditions, suggesting it could have included some content that viewers found disagreeable or confusing.

## GENERAL DISCUSSION

In the current article, we present two experiments designed to test the efficacy of three technique-based inoculation videos (Study 1) and three logic-based inoculation videos (Study 2) at reducing susceptibility to relevant manipulative content. In Study 1, all three videos increased manipulateness assessments of their relevant content compared to the control group. Similar findings were discovered in Study 2, wherein manipulateness assessments of relevant fallacious content were increased in all conditions compared to the control group. Furthermore, discernment between polarizing and neutral content was significantly increased after watching the polarization inoculation video compared to the control group in Study 1, and fallacy discernment of straw man content (versus neutral content) was significantly higher after viewing the straw man inoculation video compared to the control group in Study 2.

However, there were some inconsistencies regarding the efficacy of each video in improving the discernment of relevant content in both studies. In Study 1, discernment was not significantly better than that of the control group for the conspiracy theories or fake experts inoculation videos (but see linear mixed effects models for more promising findings). Similarly, in Study 2, discernment was not significantly higher for the whataboutism or moving the goalposts inoculation videos compared to the control group. Furthermore, while none of the inoculation videos in Study 1 significantly altered participants' abilities to detect relevant non-manipulative content compared to the control group, participants who viewed the fake experts video reported higher manipulateness assessments of irrelevant non-manipulative content than the control group,

and participants who viewed the moving the goalposts inoculation video perceived the relevant non-fallacious content as more manipulative than did those in the control group.

Mixed evidence was also obtained for our secondary hypotheses to examine the effects of our videos on confidence in detecting relevant content, motivations to resist misinformation, and sharing decisions. Only the conspiracy theories and straw man fallacy inoculation videos increased confidence in manipulateness assessments of relevant content, and even then, this only appeared consistent for manipulative but not non-manipulative content (see Supplement, Sections 13 and 14). Furthermore, while the motivation to resist misinformation was increased by all three of the logic-based inoculation videos in Study 2, the motivation to resist misinformation was not increased by the videos in Study 1 and was in fact lower in the conspiracy theories condition compared to the control group (see Supplement, Section 11). Finally, across both studies, only the straw man fallacy inoculation video increased sharing discernment of straw man fallacy and whataboutism posts in Study 2 (see Supplement, Section 15).

Taken together, these findings suggest that five of the six total videos are at least likely to increase viewers' abilities to detect relevant misinformation without making them more distrustful or gullible (see Table 1). The polarization and straw man fallacy videos are likely to result in the greatest increase in the detection of relevant manipulation considering their additional improvements in discernment, and the straw man fallacy video is also likely to increase participants' sharing decisions (see Supplement, Section 15). Nevertheless, the higher distrust of non-fallacious content after viewing the moving the goalposts inoculation video is interesting. While skepticism itself is not necessarily problematic, distrust of reliable information may push viewers into believing information from alternative unverifiable sources, such as those presented in conspiracy theories (see Douglas & Sutton, 2023; Pierre, 2020). Importantly, this could be accounted for by skepticism around unrelated content, meaning that when people are inoculated against one tactic, they may be more prone to finding false positives in unrelated non-manipulative content. With regard to the non-significant effect of the conspiracy theories video on manipulation discernment of relevant content, this may in part be explained by recent findings suggesting that inoculation only improves the detection of novel implausible conspiracy theories, but not novel plausible ones (O'Mahony et al., 2024).

The detection of logical fallacies is no easy task. de Wijze (2003) argues—in reference to the *ad hominem* fallacy—that fallacies in the real world are not as clearly detectable as those presented in textbook definitions, and it may be particularly difficult to find any argument at all that is completely devoid of fallacious reasoning. In a similar way, arguments in the real world, or indeed in the content presented in Study 2, may contain multiple fallacies at once. For example, some posts could have been perceived as moving the goalposts in order to straw man an argument or achieve whataboutism. Therefore, we warrant restraint when taking the somewhat inconsistent findings for the fallacious content as an indictment of the intervention strategies presented here. Instead, we suggest that these slightly complex findings represent issues that are present in all approaches to education. That is, while it may be easy to teach people a formula for understanding the world, the application of this formula into everyday experiences is less straightforward due to confounds and overlapping concepts. Future research should investigate these effects with variations of different content for participants to rate after watching the videos to determine the replicability of these findings.

Another ambiguity remains with the exact nature of item rating tasks as the be-all and end-all metric of inoculation intervention efficacy (see Roozenbeek et al., 2024). For example, most item rating tasks used to measure misinformation susceptibility in the literature—including the ones we present here—are unstandardized. This means that insights into the exact nature of their validity, reliability, and applicability to different contexts tends to be lacking (see Maertens, Roozenbeek, et al., 2025). In the current work, there were items intended to measure perceptions of non-manipulative content that could have understandably been perceived as manipulative (e.g., “Oil company Exxon discovered the damage that fossil fuels would have



TABLE 1 Results for the detection of relevant content in Studies 1 and 2.

Study	Tactic	Hypothesis	Result
Study 1	Polarization	Manipulation discernment	Improved
		Manipulative content	Improved
		Non-manipulative content	Null
	Conspiracy theories	Manipulation discernment	Null
		Manipulative content	Improved
		Non-manipulative content	Null
	Fake experts	Manipulation discernment	Null
		Manipulative content	Improved
		Non-manipulative content	Null
Study 2	Whataboutism	Fallacy discernment	Null
		Fallacious content	Improved
		Non-fallacious content	Null
	Moving the Goalposts	Fallacy discernment	Null
		Fallacious content	Improved
		Non-fallacious content	Backfired
	Straw Man	Fallacy discernment	Improved
		Fallacious content	Improved
		Non-fallacious content	Null

Note: Improved, hypothesis supported; Null, non-significant effect; Backfire, opposite to hypothesized effect (i.e., perceptions of non-fallacious content rated as more fallacious than in the control condition).

on the planet in the 1950s, but funded media narratives to distract from this” may be a true conspiracy theory, but a conspiracy theory nonetheless; “Economic vulnerability and right-wing extremism often go hand-in-hand: UN report” may be perceived as polarizing despite its reference to reliable research findings).

Regardless of these complexities, we posit that the use of video-based educative content to increase people's abilities to understand the concepts presented is an uncontroversial assumption. In other words, the inconsistencies detected by sensitive item rating tasks, while important for developing our theoretical understanding of inoculation interventions, which already hold a great deal of convincing evidence for their efficacy (e.g., Lu et al., 2023), could be considered secondary to our intention to provide scalable content. While people may not necessarily seek out gamified or video inoculation content themselves, prebunking videos could at least be promoted in campaigns that increase viewership without requiring viewers to seek out content or actively engage as much as they do with prebunking games (e.g., as adverts during YouTube videos; see Roozenbeek et al., 2022).

Limitations and future directions

Study 1 comprised UK participants, whereas Study 2 comprised US participants. This decision was made to address the accents of the narrators in the videos but may have reduced our ability to analyze the consistency in the findings for samples with the same national origin. Harjani et al. (2023) recently showed that inoculation messages may not work in all national contexts. Future research should thus validate the efficacy of these videos in improving discernment in other English-speaking nations (Parihar et al., 2025).



Another limitation pertains to the longevity of these interventions. Capewell et al. (2024) discovered that the effects of inoculation at reducing misinformation susceptibility decay rapidly without the immediate implementation of an item rating task. Therefore, the efficacy of the videos after a 24-hour period may decline when distributed into the real world without the item rating task included in our experiments. That said, Maertens, Roozenbeek, et al. (2025) showed that if participants were reminded of the content of the inoculation messages with booster messages, this is conducive to the long-term effectiveness of their content. The videos presented here have re-watchability that certain inoculation messages do not. Furthermore, the videos developed in the current article could be presented as YouTube adverts during a selected video so that a short survey item can be presented after the advert with the use of Google Brand Lifts (see also Roozenbeek et al., 2022), mirroring a similar approach to those in the item rating task implemented here. Future work could also test the effectiveness of different video content in prebunking the same manipulation technique to test the variability of manipulation discernment conferred by prebunking video content. While this content is expensive to create, the use of AI-generated content could be a scalable option (e.g., Linegar et al., 2024).

In one part of the experimental design implemented here, participants were being paid to take part in an experiment that teaches them how to detect (non-)manipulative content. While this could suggest that any effects are inflated due to incentivization (e.g., Panizza et al., 2022), we argue that this could also make participants more likely to spot manipulation when it is not present. That is, the active motivation to detect manipulation is not always felt in the real world. Therefore, in this experimental paradigm, perceptions of non-manipulative content as manipulative could be due to a *false-positive* bias, similar to an issue raised by Modirrousta-Galian and Higham (2023).

## CONCLUSION

We conducted two experiments testing the efficacy of three technique-based (Study 1) and three logic-based (Study 2) inoculation videos at improving manipulation and fallacy discernment. While findings were somewhat complex, the increased detection of relevant manipulative content was increased after viewing all videos compared to the control group. We posit that while important to consider, the inconsistencies in these findings are secondary to providing scalable inoculation videos, considering the existing evidence of the efficacy of inoculation interventions (see Lu et al., 2023). We urge researchers to run similar experiments to further validate the efficacy of these videos, presenting participants with items that have been validated more than the ones included here. We hope these findings can be used to support the distribution of these videos in campaigns that circumvent the issue of viewers being unlikely to seek out this content themselves (e.g., as un-skippable adverts before YouTube videos are played).

## ACKNOWLEDGMENTS

We would like to give special thanks to Luke Newbold and Sean Sears at Studio You, as well as the Lens Change team for their responsive and creative work developing the prebunking videos.

## FUNDING INFORMATION

This research was supported by the Infodemic Grant from the UK Cabinet Office. We wish to acknowledge funding from the Department of Foreign Affairs and Trade Ireland.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in OSF at <https://osf.io/8g7jc/>.

## ORCID

Mikey Biddlestone  <https://orcid.org/0000-0003-1438-7392>

Jon Roozenbeek  <https://orcid.org/0000-0002-8150-9305>

## REFERENCES

- Allen, J., Howland, B., Mobius, M. M., Rothschild, D. M., & Watts, D. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14), eadk3539. <https://doi.org/10.1126/sciadv.aay3539>
- Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, 384(6699), eadk3451. <https://doi.org/10.1126/science.adk3451>
- Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, 39(2), 184–207. <https://doi.org/10.1111/hcre.12000>
- Banas, J. A., & Richards, A. S. (2017). Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. *Communication Monographs*, 84(2), 164–178. <https://doi.org/10.1080/03637751.2017.1307999>
- Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. V. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, 8(1). <https://doi.org/10.1177/20539517211013868>
- Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, 3(1), 2. <https://doi.org/10.5334/joc.91>
- Biddlestone, M., Azevedo, F., & van der Linden, S. (2022). Climate of conspiracy: A meta-analysis of the consequences of belief in conspiracy theories about climate change. *Current Opinion in Psychology*, 46, 101390. <https://doi.org/10.1016/j.copsyc.2022.101390>
- Biddlestone, M., Green, R., Douglas, K. M., Azevedo, F., Sutton, R. M., & Cichocka, A. (2025). Reasons to believe: A systematic review and meta-analytic synthesis of the motives associated with conspiracy beliefs. *Psychological Bulletin*, 151(1), 48–87. <https://doi.org/10.1037/bul0000463>
- Biddlestone, M., Roozenbeek, J., & van der Linden, S. (2022). Once (but not twice) upon a time: Narrative inoculation against conjunction errors indirectly reduces conspiracy beliefs and improves truth discernment. *Applied Cognitive Psychology*, 37(2), 304–318. <https://doi.org/10.1002/acp.4025>
- Boudry, M. (2022). Why we should be suspicious of conspiracy theories: A novel demarcation problem. *Episteme*, 20(3), 611–631. <https://doi.org/10.1017/epi.2022.34>
- Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. *Frontiers in Psychology*, 4, 225. <https://doi.org/10.3389/fpsyg.2013.00225>
- Capewell, G., Maertens, R., Remshard, M., Compton, J., van der Linden, S., Lewandowsky, S., & Roozenbeek, J. (2024). Misinformation interventions decay rapidly without an immediate post-test. *Journal of Applied Social Psychology*, 54(8), 441–454. <https://doi.org/10.1111/jasp.13049>
- Compton, J., & Ivanov, B. (2012). Untangling threat during inoculation-conferred resistance to influence. *Communication Reports*, 25(1), 1–13. <https://doi.org/10.1080/08934215.2012.661018>
- Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One*, 12(5), e0175799. <https://doi.org/10.1371/journal.pone.0175799>
- de Wijze, S. (2003). Complexity, relevance and character: Problems with teaching the ad hominem fallacy. *Educational Philosophy and Theory*, 35(1), 31–56. <https://doi.org/10.1111/1469-5812.00004>
- Douglas, K. M., & Sutton, R. M. (2023). What are conspiracy theories? A definitional approach to their correlates, consequences, and communication. *Annual Review of Psychology*, 74(1), 271–298. <https://doi.org/10.1146/annurev-psych-032420-031329>
- Edwards, T., Giannelis, A., Willoughby, E. A., & Lee, J. J. (2024). Predicting political beliefs with polygenic scores for cognitive performance and educational attainment. *Intelligence*, 104, 101831. <https://doi.org/10.1016/j.intell.2024.101831>
- Green, M., McShane, C. J., & Swinbourne, A. (2022). Active versus passive: Evaluating the effectiveness of inoculation techniques in relation to misinformation about climate change. *Australian Journal of Psychology*, 74(1), 2113340. <https://doi.org/10.1080/00049530.2022.2113340>
- Hanel, P. H., Maio, G. R., & Manstead, A. S. (2019). A new way to look at the data: Similarities between groups of people are large and important. *Journal of Personality and Social Psychology*, 116(4), 541–562. <https://doi.org/10.1037/pspi0000154>
- Harjani, T., Basol, M., Roozenbeek, J., & van der Linden, S. (2023). Gamified inoculation against misinformation in India: A randomized control trial. *Journal of Trial and Error*, 3(1), 14–56. <https://doi.org/10.36850/e12>
- Hruschka, T. M., & Appel, M. (2023). Learning about informal fallacies and the detection of fake news: An experimental intervention. *PLoS One*, 18(3), e0283238. <https://doi.org/10.1371/journal.pone.0283238>

- Ivanov, B., McVicker, S. M., & Gordon, M. (2023). Propaganda education through exposure to authentic content. *The Journal of Communication and Media Studies*, 9(1), 45–62. <https://doi.org/10.18848/2470-9247/cgp/v09i01/45-62>
- Jaeger, P. T., & Taylor, N. G. (2021). Arsenals of lifelong information literacy: Educating users to navigate political and current events information in world of ever-evolving misinformation. *The Library Quarterly*, 91(1), 19–31. <https://doi.org/10.1086/711632>
- Jigsaw. (2023, October 25). Prebunking to build defenses against online manipulation tactics in Germany. *Medium*. <https://medium.com/jigsaw/prebunking-to-build-defenses-against-online-manipulation-tactics-in-germany-aldbfbc67ala>
- Kuru, O. (2024). Literacy training vs. psychological inoculation? Explicating and comparing the effects of predominantly informational and predominantly motivational interventions on the processing of health statistics. *Journal of Communication*, 75(1), 64–78. <https://doi.org/10.1093/joc/fjae032>
- Lessenski, M. (2023). *Bye bye Birdie: The challenges of disinformation*. Media and Learning Association. <https://media-and-learning.eu/type/featured-articles/bye-bye-birdie-the-challenges-of-disinformation/>
- Lewandowsky, S. (2024). Truth and democracy in an era of misinformation. *Science*, 386(6717), eads5695. <https://doi.org/10.1126/science.ads5695>
- Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353–369. <https://doi.org/10.1016/j.jarmac.2017.07.008>
- Lewandowsky, S., Ecker, U. K., Cook, J., van der Linden, S., Roozenbeek, J., & Oreskes, N. (2023). Misinformation and the epistemic integrity of democracy. *Current Opinion in Psychology*, 54, 101711. <https://doi.org/10.1016/j.copsyc.2023.101711>
- Lewandowsky, S., Ecker, U. K., Cook, J., van der Linden, S., Roozenbeek, J., Oreskes, N., & McIntyre, L. C. (2024). Liars know they are lying: Differentiating disinformation from disagreement. *Humanities and Social Sciences Communications*, 11(1), 1–14. <https://doi.org/10.1057/s41599-024-03503-6>
- Lewandowsky, S., & Yesilada, M. (2021). Inoculating against the spread of Islamophobic and radical-islamist disinformation. *Cognitive Research: Principles and Implications*, 6(1), 57. <https://doi.org/10.1186/s41235-021-00323-z>
- Linegar, M., Sinclair, B., van der Linden, S., & Alvarez, R. M. (2024). Prebunking elections rumors: Artificial intelligence assisted interventions increase confidence in American elections. *ArXiv preprint ArXiv: 2410.19202*. <https://arxiv.org/abs/2410.19202>
- Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X. (2023). Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 25, e49255. <https://doi.org/10.2196/49255>
- Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & van der Linden, S. (2023). The misinformation susceptibility test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, 56(3), 1863–1899. <https://doi.org/10.3758/s13428-023-02124-2>
- Maertens, R., Roozenbeek, J., Simons, J., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., & van der Linden, S. (2025). Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nature Communications*, 16(1), 2062. <https://doi.org/10.1038/s41467-025-57205-x>
- McGuire, W. J. (1961). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Sociometry*, 24(2), 184. <https://doi.org/10.2307/2786067>
- McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology*, 62(2), 327–337. <https://doi.org/10.1037/h0042026>
- Miller, C. H., Ivanov, B., Sims, J., Compton, J., Harrison, K. J., Parker, K. A., Parker, J. L., & Averbek, J. M. (2013). Boosting the potency of resistance: Combining the motivational forces of inoculation and psychological reactance. *Human Communication Research*, 39(1), 127–155. <https://doi.org/10.1111/j.1468-2958.2012.01438.x>
- Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*, 152(9), 2411–2437. <https://doi.org/10.1037/xge0001395>
- Neylan, J., Biddlestone, M., Roozenbeek, J., & van der Linden, S. (2023). How to “inoculate” against multimodal misinformation: A conceptual replication of Roozenbeek and van Der Linden (2020). *Scientific Reports*, 13(1), 18273. <https://doi.org/10.1038/s41598-023-43885-2>
- O'Mahony, C., Murphy, G., & Linehan, C. (2024). True discernment or blind scepticism? Comparing the effectiveness of four conspiracy belief interventions. *Advances in Psychology*, 2(1), 1–28. <https://doi.org/10.56296/aip00030>
- Panizza, F., Ronzani, P., Martini, C., Mattavelli, S., Morisseau, T., & Motterlini, M. (2022). Lateral reading and monetary incentives to spot disinformation about science. *Scientific Reports*, 12(1), 5678. <https://doi.org/10.1038/s41598-022-09168-y>
- Parihar, S., Harjani, T., Mathur, P., Goldberg, B., van der Linden, S., & Roozenbeek, J. (2025). Countering misinformation in India through prebunking. *PsyArXiv Preprints*. [https://doi.org/10.31234/osf.io/g2y9z\\_v2](https://doi.org/10.31234/osf.io/g2y9z_v2)

- Pierre, J. M. (2020). Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology*, 8(2), 617–641. <https://doi.org/10.5964/jspp.v8i2.1362>
- Roozenbeek, J., Remshard, M., & Korychenko, Y. (2024). Beyond the headlines: On the efficacy and effectiveness of misinformation interventions. *Advances in Psychology*, 2(1), 1–17. <https://doi.org/10.56296/aip00019>
- Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1), 65. <https://doi.org/10.1057/s41599-019-0279-9>
- Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34), eabo6254. <https://doi.org/10.1126/sciadv.abo6254>
- Tartaglione, J., & De-Wit, L. (2024). How the manner in which data is visualized affects and corrects (mis)perceptions of political polarization. *British Journal of Social Psychology*, 64(1), e12787. <https://doi.org/10.1111/bjso.12787>
- Tindale, C. W. (2007). *Fallacies and argument appraisal*. Cambridge University Press.
- van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3), 460–467. <https://doi.org/10.1038/s41591-022-01713-6>
- van der Linden, S. (2023). Psychological inoculation against misinformation. *Journal of Neurology, Neurosurgery & Psychiatry*, 94(12), e2.40. <https://doi.org/10.1136/jnnp-2023-bnpa.9>
- van der Linden, S., Albarracin, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2023). Using psychological science to understand and fight health misinformation: An APA consensus statement. *PsycEXTRA Dataset*. <https://doi.org/10.1037/e506432023-001>
- van der Linden, S., & Korychenko, Y. (2024). A broader view of misinformation reveals potential for intervention. *Science*, 384(6699), 959–960. <https://doi.org/10.1126/science.adp9117>
- van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2), 1600008. <https://doi.org/10.1002/gch2.201600008>
- van der Linden, S., & Roozenbeek, J. (2024). “Inoculation” to resist misinformation. *JAMA*, 331(22), 1961–1962. <https://doi.org/10.1001/jama.2024.5026>
- van Prooijen, J., & Douglas, K. M. (2017). Conspiracy theories as part of history: The role of societal crisis situations. *Memory Studies*, 10(3), 323–333. <https://doi.org/10.1177/1750698017701615>
- Walton, D. N. (1995). *A pragmatic theory of fallacy*. University of Alabama Press.
- World Economic Forum. (2024, January 10). *Global risks 2024: Disinformation tops global risks 2024 as environment threats intensify*. World Economic Forum. <https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Biddlestone, M., Roozenbeek, J., Suiter, J., Culloty, E., & van der Linden, S. (2025). Tune in to the prebunking network! Development and validation of six inoculation videos that prebunk manipulation tactics and logical fallacies in misinformation. *Political Psychology*, 46, 1858–1886. <https://doi.org/10.1111/pops.70015>