



Countering misinformation through psychological inoculation

Sander van der Linden*

Department of Psychology, University of Cambridge, Cambridge, United Kingdom

*Corresponding author. e-mail address: sander.vanderlinden@psychol.cam.ac.uk

Contents

1. Introduction	2
1.1 Defining misinformation	3
1.2 From debunking to prebunking	4
2. Inoculation theory: Origins	5
3. Inoculation against misinformation (about climate change)	8
3.1 Therapeutic versus prophylactic inoculation	10
3.2 Narrow-spectrum versus generalized (broad-spectrum) immunity	12
4. Game on: Active versus passive inoculation	13
4.1 Bad news	13
4.2 Good news about bad news	16
4.3 Item and testing effects	20
4.4 GoViral, harmony square, and other inoculation games	23
4.5 Scaling inoculation	26
5. Mechanisms: The memory-motivation model of inoculation theory	29
6. Limitations and future directions	33
6.1 Cross-cultural research	33
6.2 Prebunking versus inoculation	36
6.3 Side effects: Skepticism, real news, and discernment	37
6.4 Psychological herd immunity	45
7. Conclusion	48
Acknowledgments	49
References	49

Abstract

The global spread of misinformation is one of the defining challenges of our era. This chapter reviews findings from an ongoing research program that examines the following key question: can people become inoculated against misinformation? After a brief overview of early research into inoculation theory, I discuss critical theoretical advancements and review evidence from both the lab and field that people can indeed gain relative immunity to misinformation across a wide range of contexts, from public health

to political elections. I will discuss empirical evidence for the key mechanisms that help explain why psychological inoculation is effective: (1) forewarning people and (2) prebunking misinformation via a weakened dose, or equipping people with the skills they need—in advance—to refute future falsehoods. Finally, I will discuss how the analogy has been extended to include diffusion of the “vaccine” in social networks, and the possibility of herd immunity. Limitations and avenues for future research are discussed.



1. Introduction

The unprecedented spread of misinformation poses a threat to science, public health, elections, and the functioning of our democracies (Lewandowsky et al., 2017; van der Linden, 2023a; Swire-Thompson & Lazer, 2019; West & Bergstrom, 2021). The World Health Organization (WHO) declared a worldwide “infodemic” in 2020, which they define as an overabundance of information, especially misinformation (Zarocostas, 2020). Although clearly not a new phenomenon, in 2022, the World Economic Forum noted that “misinformation” is one of the areas with the least established international risk mitigation efforts (WEF, 2022). Misinformation has shown to impact people’s support for important societal issues such as climate change (van der Linden, Maibach et al., 2017; van der Linden, Leiserowitz et al., 2017), their willingness to comply with public health guidance (Roosenbeek et al., 2020, 2021), decisions whether or not to get vaccinated (Loomba et al., 2021; Pierri et al., 2022; Wilson & Wiysonge, 2020), voting in elections (Gunther, Beck, & Nisbet, 2019), as well as support for political violence, including vandalism, injury, and in some cases even death (Jolley & Paterson, 2020; Goel et al., 2018; Hebel-Sela, Hameiri, & Halperin, 2022).

Models from epidemiology are increasingly used to study the diffusion of misinformation in social networks, finding that (mis)information indeed spreads much like a virus (Bak-Coleman et al., 2022; Jiang, Gao, & Zhuang, 2021; van der Linden, 2022). The question which naturally arises is whether it is then also possible to “immunize” or “inoculate” people against misinformation? In this chapter, I provide an overview of an on-going research program which has investigated the question of when and how people can become relatively more immune to misinformation through prebunking and psychological inoculation. In fact, it is worth noting that the last entry in *Advances in Experimental Social Psychology* on this topic was published exactly 60 years ago by William J. McGuire (McGuire, 1964). Research on inoculation theory has been revived in the context of countering misinformation in the digital era and I will summarize not only what we have learned both in the

lab and in the field but also delineate new theoretical advancements, some critiques and limitations, as well as key avenues for future research.

1.1 Defining misinformation

Before I proceed, it is important to note that the definition that scholars in different fields use for the term “misinformation” can greatly impact not only research study and stimuli designs but also the scope of the problem. For example, misinformation based on narrow definitions that are restricted to only include content that is utterly “fake” or entirely fabricated (e.g., flat earth) is not very common. On the other hand, broader definitions of misinformation—which include biased or misleading news—cover a much wider problem. Partly for that reason, misinformation prevalence estimates in our media ecosystems can vary widely, from 0.2% to 29% of people’s media diet (Altay et al., 2022; van der Linden et al., 2023).

Some scholars define “fakeness” at the level of the source of an article rather than the content because low-credibility outlets, on average, are often more likely to share misinformation than highly credible outlets (see Altay et al., 2022; Grinberg et al., 2019). Another popular approach is to select headlines based on whether they have been fact-checked or not (e.g., Pennycook & Rand, 2019) or to benchmark claims against the current scientific consensus (e.g., Vraga & Bode, 2020; van der Linden, Leiserowitz et al., 2017). All of these approaches are valid but have limitations. For example, misinformation is sometimes published or repeated by credible outlets which can have much greater impact on public opinion (Goel et al., 2023; Traberg, 2022); scientific consensus is not always available especially when emerging science is subject to uncertainty and revision (Swire-Thompson & Lazer, 2019); and most misinformation is not fact-checked and usually not completely false but rather manipulative or misleading in some way (Wardle, 2018). To give a concrete example, the following headline was published by the *Chicago Tribune* (a credible outlet) in 2021: *A ‘healthy’ doctor died two weeks after getting a COVID-19 vaccine; CDC is investigating why*. This headline was not only published by a credible outlet, but the claim itself is also not technically false. After all, a healthy doctor did die two weeks after getting the Covid-19 vaccine. Yet, the headline is framed in such a way to falsely imply a causal connection (where there is only correlation) and thereby plays into anti-vaccination sentiment. In fact, it became the most shared story on Facebook in 2021, especially among anti-vaccination groups (Benton, 2021) and this type of misinformation has shown to be much more damaging to vaccination attitudes (Allen et al., 2023).

Accordingly, our lab has focused mostly on defining misinformation not in terms of “true” or “fake” but rather in terms of the presence of common misinformation techniques, which is an important distinction and very close to the definition adopted by the APA Consensus Report on the Psychology of Misinformation: “*any information that is demonstrably false or misleading, regardless of its source or intention*” (van der Linden et al., 2023).

1.2 From debunking to prebunking

Psychologists have long studied how to correct false information and ways to counter propaganda (e.g., Osborn, 1939; Lumsdaine & Janis, 1953; Greene, Flynn, & Loftus, 1982), and many meta-analyses now exist on the efficacy of fact-checking (e.g., Walter & Tukachinsky, 2019) and debunking falsehoods (Chan et al., 2017). Although there is increasing international evidence that debunking and fact-checking can at least be partially effective (Porter & Wood, 2021; Walter et al., 2020), a recent meta-analysis found no significant average effect of corrections in the context of scientific misinformation (Chan & Albarracín, 2023). It should be noted that it is well-known that debunking can be more or less effective depending on the issue domain (e.g., health vs politics), the strength of an individual’s pre-existing beliefs and ideologies and the depth and quality of the fact-check in question (see Bruns et al., 2023; Chan et al., 2017; Walter & Murphy, 2018). Importantly, though, several major challenges remain when trying to correct misinformation after it has already spread.

The first concerns the so-called “backfire effect” where a correction ironically strengthens people’s original belief in the myth (Nyhan & Reifler, 2010). Backfire or “boomerang” effects can occur for a variety of reasons (Lewandowsky et al., 2012). For example, the fact-check may not be congenial to a person’s ideological worldview (“*worldview* backfire”) or it may repeat the myth too prominently relative to the correction so that a debunking attempt inadvertently strengthens people’s familiarity with the myth (“*familiarity* backfire”). The emerging consensus is that concerns about “backfire” have been exaggerated in both the social scientific literature and the media for many years (Lewandowsky et al., 2020; Nyhan, 2021) given that both the *generalizability* and the *prevalence* of backfire effects have been called into question (Swire-Thompson, DeGutis, & Lazer, 2020; Wood & Porter, 2019). Nevertheless, it is fair to say that backfire effects *can* occur especially among non-receptive audiences. For example, a recent study noted that debunking misinformation about Covid-19 vaccines (“mRNA vaccines do not contain live virus”) corrected the misperception but made people unintentionally more worried about live vaccines, especially if they were already vaccine hesitant (Krause et al., 2023).

The second issue is that while misinformation often goes viral, factual information typically does not (Vosoughi, Roy, & Aral, 2018) and so from an applied perspective, fact-checks are always running behind the curve (Roozenbeek & van der Linden, 2020). One famous example concerns the 2017 court-ordered corrective ads from the tobacco industry, which had to inform the general public that they had deliberately lied about the link between smoking and health risks for many decades. The corrective ads received little to no engagement on social media (Kostygina et al., 2020) and only reached an estimated 40% of the population (Blake, Willis, & Kaufman, 2020; Chido-Amajuoyi et al., 2019). The prevalence of tobacco-related misinformation is unknown but false information has been advertised by the tobacco industry for 50 years and studies found that repeated false claims about tobacco are more influential than repeated true claims (Morgan & Cappella, 2023). Moreover, the corrective ads only constituted about 0.6% of the industry's total marketing budget. Perhaps unsurprisingly, misperceptions about the health risks of smoking have been widespread (Cummings et al., 2002; Frieden & Blakeman, 2005).

The final problem concerns the robust finding that, once people have been exposed to misinformation, they continue to retrieve and rely on false details from memory when asked to make inferences about an event despite having seen a correction, a finding known as the *continued influence of misinformation* (Ecker et al., 2022; Lewandowsky et al., 2012; Seifert, 2002). Popular examples include the persistent myth that vaccines cause autism (Motta & Stecula, 2021) or that Iraq harbored Weapons of Mass Destruction (Lewandowsky et al., 2005). Meta-analyses have repeatedly confirmed that corrections reduce but do not fully eliminate belief in misinformation (Chan et al., 2017; Walter & Tukachinsky, 2019).

Perhaps somewhat unsurprisingly then, studies and systematic reviews often (but not exclusively) find that prevention is better than cure when it comes to misinformation (see Jolley & Douglas, 2017; O'Mahony et al., 2023), or to *prebunk* rather than only debunk. But what do we mean exactly by “prebunking” and psychological “inoculation”?



2. Inoculation theory: Origins

Although McGuire (together with his then doctoral student Demetrios Papageorgis) is rightly credited with initially developing the theory of psychological inoculation as we know it in social psychology

today (McGuire & Papageorgis, 1961), he certainly wasn't the first to think of the idea. Indeed, about 20 years prior, educational psychologists were already wondering if it would be possible to "immunize" students against propaganda in the classroom (see Osborn, 1939). Perhaps the more fundamental idea of refuting an opponent's misleading arguments in advance can be traced all the way back to Aristotle as described in his infamous *Rhetoric* (Compton, 2005; Compton & Pfau, 2005; van der Linden, 2023a). But McGuire *was* certainly the first to explain a finding Irving Janis could only speculate on. In the early 50's, Lumsdaine and Janis (1953) conducted an experiment where they exposed students to either a one-sided or a two-sided message about the Soviet Union's nuclear capabilities. The goal was to convince students that Russia did not have the capacity to produce nuclear bombs in the next few years, after which the students were exposed to "counterpropaganda" arguing the exact opposite. What they found was interesting, namely that the students in the two-sided message condition formed stronger resistance to the countermesssage. Lumsdaine and Janis (1953) hypothesized that raising and refuting some elements of the countermesssage in advance (the "two-sided" message condition) effectively "inoculated" participants. McGuire (1961a) and McGuire and Papageorgis (1961) later formalized these findings following a series of experiments. The basic gist of inoculation theory is that, just as vaccines expose people to a weakened dose of a pathogen, triggering the body's immune system to help confer resistance against future infection, by providing people with a severely weakened dose of a persuasive argument—and by strongly refuting that argument in advance—people can build up cognitive resistance to persuasion. Or in McGuire's own words: *We can develop belief resistance in people as we develop disease resistance in a biologically overprotected man or animal: by exposing the person to a weak dose of the attacking material, strong enough to stimulate his [or her] defenses, but not strong enough to overwhelm them* (McGuire, 1970, p. 37).

On a theoretical level, McGuire's research program was motivated by two related principles. The first was the limited evidence for the efficacy of one-sided or "supportive" messages in conferring resistance to persuasion (McGuire, 1961a). Following the Korean War, the U.S. government was concerned about the apparent brainwashing of U.S. soldiers by enemy ("communist") forces. The government's response was that the soldiers simply needed more facts and supportive information about American values, but McGuire disagreed. Facts were not the issue. He hypothesized that the soldiers had never been exposed to an attack on Western values

and ideologies, so they had no mental defenses at the ready. Instead, they needed a weakened dose of the types of attacks they might be facing on their belief system and practice resisting these attacks with persuasive counterarguments (McGuire & Papageorgis, 1961). But McGuire did not think the Korean War a good test case for the lab because his students were divided on the communism versus capitalism question. This insight led to the second principle: selective exposure, or the tendency for people to avoid exposing themselves to viewpoints that challenge their existing attitudes. McGuire (1961, 1970) was relatively unconvinced that people generally avoid evidence that challenges existing beliefs, so he wanted to constrain his initial experiments to belief situations where he could more or less guarantee that people had never been exposed to counterevidence before. A clean test case for inoculation would therefore involve what he referred to as “cultural truisms” or ideas so widely held that nobody would really question them. These ideas subject to selective exposure essentially live in a “germ-free” ideological environment, i.e., they are strong and salient but vulnerable when exposed to hostile material (McGuire, 1970).

The basic idea is thus that people (1) already need to have a favorable attitude toward the target issue and (2) not been exposed to counter-attitudinal arguments before. In a typical experiment, McGuire would test students on an issue such as the health benefits of brushing your teeth or getting regular medical check-ups. The logic being that most students would support these arguments and probably have never come across serious challenges to those beliefs. In multiple experiments, McGuire found that when compared to one-sided supportive messages that merely reinforced the “truism”, exposing students to a “refutational pre-emption” which contained a weakened dose of the counterarguments (e.g., that regular check-ups turn people into hypochondriacs), produced greater resistance to the full countermessage later on (McGuire, 1964). Importantly, to reiterate, the inoculation condition focused mostly on raising and refuting the counterclaims (not supporting facts for the truism). In the control condition, participants saw no message as it served as a baseline measure for belief in the claim, which was usually measured pre and post exposure on a 15-point scale (from definitely false to definitely true). Whereas the supportive messages had a marginal positive effect on the students’ ratings of the claims, the inoculation produced much greater resistance to future persuasion. McGuire went on to test variations on these experiments but ultimately dropped this research program after his *Advances in Experimental Social Psychology* chapter on attitudes and resistance to

persuasion (McGuire, 1964). Although some related scholarship was maintained over the years by others, Eagly and Chaiken (1993) summarize in their landmark text on the psychology of attitudes that “*although the analogy is admittedly clever and valid the theory has not seen much development for many years and many of the questions it raised remain unresolved*” (p. 568). The biggest open question was perhaps posed by McGuire himself (McGuire, 1964, p. 1) and concerns the extent to which inoculation generalizes and can predict immunity to persuasion in the case of controversial beliefs.



3. Inoculation against misinformation (about climate change)

Because inoculation theory had somewhat ironically never been tested in the context of misinformation or propaganda (the original real-world concern that motivated development of the theory), we wanted to start by doing the opposite of what McGuire had worked on. Instead of a “germ-fee” ideological environment we wanted to select an issue for which people would have had clear differential prior media exposure. In fact, we wanted to select a real-world polarizing issue characterized by the presence of misinformation and propaganda. Given concerted disinformation campaigns seeking to sow doubt about the scientific consensus on human-caused climate change (Oreskes & Conway, 2010; van der Linden, 2023a), and the fact that belief in climate change is highly divided along partisan lines in the US (Ballew et al., 2019), it seemed like the perfect test case. Can inoculation theory confer resistance to misinformation about a real-world issue? In the first set of lab experiments, we followed McGuire’s (1970) original paradigm closely. We first conducted a national survey to establish which types of misinformation the American public were most familiar with, which reinforced the hypothesis that casting doubt on the scientific consensus was one of the most persuasive misinformation claims. We then recruited 2167 Americans from *Mturk* and randomly assigned them into one of five experimental conditions: (1) a *facts-only* condition where participants were simply informed of the facts or that *97% of climate scientists agree that human-caused climate change is real* (akin to McGuire’s “supportive” or “one-sided message”), a (2) *misinformation-only* condition where participants were exposed to a website hosting a real-world petition which claims that over 30,000 scientists agree that anthropogenic climate change is not real (known as the *Oregon Global Warming Petition*)

which went viral on social media in 2016 (Readfearn, 2016), (3) a *false balance* condition where participants saw both the scientific consensus and the misinformation side by side (mimicking the problematic media practice of pitting individual contrarians or deniers against the scientific consensus, also known as journalistic false balance; see Koehler, 2016), (4) a *forewarning-only* inoculation where participants were simply forewarned that politically motivated actors use misleading techniques to dupe people on the issue of climate change, (5) a *detailed* inoculation condition which contained both the forewarning and a preemptive refutation or *prebunk* of the misinformation, and finally (6) a control group no-message condition to establish baseline belief. The prebunk contained a weakened dose of the claims presented by the petition which were then persuasively refuted (e.g., participants were told that signatories supposedly include members of the Spice Girls and Charles Darwin) and we deconstructed the fake expert technique employed by the petition (e.g., *calling yourself a scientist doesn't make someone an expert in climate science*).

McGuire assumed that the threat of being exposed to propaganda or counter-attitudinal views would elicit *motivation* for people to defend themselves and that the preemptive refutation would give people the actual *ability* to do so (McGuire, 1964). We split out the forewarning from the full inoculation because McGuire originally assumed that the “threatening” phase of the inoculation would be implicit in the weakened dose, but later research has shown that forewarning people about false information more explicitly can also be effective on its own (Greene, Flynn, & Loftus, 1982; Petty & Cacioppo, 1977), so this represented an important opportunity to differentiate the value of the forewarning from the prebunk specifically given on-going debates as to which mechanism is more important for generating resistance (Banas & Rains, 2010). We measured people’s perception of the scientific consensus pre and post exposure in all groups on a scale which ranged from 0% to 100% (people were asked to produce estimates on several issues to obscure the purpose of the study). They were also told that they would be randomly assigned a media topic using a faux number generator (in reality, all participants were always assigned to the climate topic).

The main effects (Fig. 1A) revealed that the facts by themselves actually had a large positive effect on how people updated their views about the scientific consensus on global warming when compared to the control condition ($d = 1.23$, $n = 338$), whereas the misinformation had a clear negative impact ($d = -0.48$, $n = 392$). Somewhat concerningly, the presence

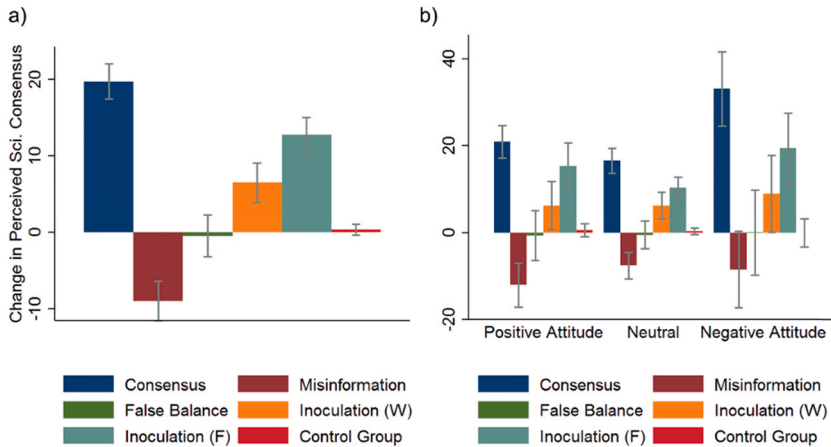


Fig. 1 The three attitudinal groups were pre-screened based on whether respondents believed that human-caused climate change is happening (“positive”) or not (“negative”) with “unsure” classified as “neutral.” The same patterns emerged when results are broken down by political party ID (Republican, Democrat, Independent). *Note:* Error bars represent 95% confidence intervals (change scores). *Inoculating against misinformation about climate change, adapted with permission from van der Linden et al. (2017).*

of misinformation completely canceled out the positive effect of facts in the false-balance condition ($d = 0.04$, $n = 352$) highlighting the malleability of public perceptions of scientific consensus (Koehler, 2016). The key question in our study of course was whether people’s attitudes about the scientific consensus could subsequently be immunized against the misinformation attack. Here we found clear significant impacts of both the forewarning-only ($d = 0.33$, $n = 363$) and the detailed inoculation conditions ($d = 0.75$, $n = 362$). From the effect-sizes it is clear that, although the warning component of inoculation proved useful in and of itself, the prebunk did most of the heavy-lifting. (Fig. 2).

3.1 Therapeutic versus prophylactic inoculation

But was the inoculation equally effective among those with differing prior attitudes toward climate change? The short answer is *yes*. We found that the inoculation worked equally well for those who believed in human-caused climate change, for those who did not believe in climate change at all, as well as for those who were on the fence about it (Fig. 1B). The same pattern of results emerged for Republicans, Democrats, and Independents. So even though these audiences likely had very different levels of prior

exposure to the myth in question, the detailed inoculation still proved effective when compared against the false balance and control conditions. But what does this imply for inoculation at a theoretical level? Is it even possible to “inoculate” those who were already exposed to and familiar with the misinformation? I do not take these results to mean that the inoculation changed the minds of climate skeptics on the issue per se (there is also more variation for the “negative” attitude group), but it did at least, on average, prevent further entrenchment.

About three years later, we conducted a pre-registered replication of this study with one crucial change: we delayed the presentation of misinformation by one week to examine if those inoculated at T_0 would still show resistance one week later, which is exactly what we found, and the inoculation proved once again to have a positive effects across the ideological spectrum (Maertens et al., 2020). Later studies using very similar stimuli and measures (e.g., free-market ideology) also replicated these general findings (e.g., see Cook et al., 2017; Cook, Ellerton, & Kinkead, 2018; Cook, van der Linden et al., 2018), including among adolescents in Austria (Schubatzky & Haagen-Schützenhöfer, 2023).¹

It turns out that other scholars have also questioned McGuire’s restriction that inoculation can only be effective on cultural truisms (see Pryor & Steinfatt, 1978; Ivanov et al., 2017; Wood, 2007). The fact the inoculation can be effective across pre-existing attitude levels has ignited a new theoretical field of inquiry where scholarship has started to distinguish between “therapeutic” and “prophylactic” inoculation campaigns (Amazeen et al., 2022; Compton, 2020; Compton et al., 2021; Traber et al., 2022). Although the ideal scenario for inoculation of course remains fully pre-emptive or prophylactic (before people are exposed to any misinformation), just as advances in medicine have started to trial and approve “therapeutic” vaccines (such as for HPV and certain types of cancer)—which can still boost immunity for people who already have the “disease”—research has shown that inoculation can still be effective even for those who may have already been exposed to the misinformation “virus” (Compton, 2020; Compton et al., 2021; Lewandowsky & van der Linden, 2021). It is perhaps helpful to

¹ Ceiling effects in public opinion on climate change and participants not believing the misinformation in the first place can present challenges to demonstrating the value of inoculation (see Williams & Bond, 2020; Schmid-Petri & Burger, 2022). More mixed findings of therapeutic inoculation (on prior attitudes) have been reported in the context of COVID-19 (see Amazeen et al., 2022; Vivion et al., 2022).

think of inoculation on a spectrum ranging from fully prophylactic to fully therapeutic depending on an individual's "exposure" status, which may or may not be known in the real-world² at the time of intervention. Although our findings largely support McGuire and Papageorgis' (1962) finding that forewarnings coupled with preemptive refutations generate greater resistance than forewarnings alone, there is some on-going debate about whether the key mechanisms underlying inoculation are the same for therapeutic versus prophylactic inoculation campaigns (Amazeen et al., 2022; Compton et al., 2021; Ivanov et al., 2022).

3.2 Narrow-spectrum versus generalized (broad-spectrum) immunity

A perhaps obvious limitation of the classical inoculation paradigm is that you can hardly inoculate people against every single myth that might present itself in the future. McGuire had in fact experimented with what he called "refutational-*same*" versus "refutational-*different*" arguments where the former raises and refutes weakened doses of the exact persuasive attack people will be exposed to in the future and the latter addresses novel arguments not raised in the subsequent attack. Cook et al. (2017) conducted a very similar study in the context of climate change (using the same misinformation treatment as our group) but with one key twist: instead of using a weakened dose of the petition itself they relied on a broader "refutational-*different*" strategy by exposing participants to an example of the fake expert technique in the context of the tobacco industry. Specifically, they warned participants that the tobacco industry had been trying to undermine the medical consensus on the health risks of smoking by using clever tactics such as the claim that "20,000 physicians say that Luckies are less irritating". Cook et al. (2017) found that by inoculating people against the fake expert technique in the context of health, they became more resistant when exposed to the same technique in the context of climate change. Interestingly, inoculation can confer so-called "cross-protection" where people become more resistant to arguments that are related but not specifically refuted in the initial inoculation (Parker, Rains, & Ivanov, 2016). This led our research group to hypothesize that *if* we can identify and inoculate people against the general techniques used in the

² At which point therapeutic inoculation devolves into debunking is conceptually not entirely clear and merits further investigation but the key practical distinction between debunking and inoculation is the format: debunking often repeats the myth and corrects it whereas inoculation forewarns and refutes a micro-dose of the misinformation.

production of misinformation, *then* people should evidence broader-scale immunity, much like a vaccine can protect people not only against specific but also against related strains of the same virus. In other words, if you can inoculate people against the basic building blocks of a conspiracy theory, they should become more immune to a whole range of (related) conspiracy theories regardless of their specific content.



4. Game on: Active versus passive inoculation

In the climate experiments we used an essay which essentially gave participants the refutations they needed—in advance—to withstand the misinformation “attack” on their attitudes much like McGuire had conducted his initial experiments with cultural truisms. This is a fairly passive process by which the experimenter essentially provides the participants with the motivation (threat) and refutations (prebunk). But McGuire wondered whether it could be more effective to let participants actively generate their own counterarguments, which he referred to as “active inoculation” (McGuire, 1961b). In terms of the vaccine analogy, we can think of it as having people generate their own mental antibodies, though this concept has remained mostly unexplored in the literature (Banas & Rains, 2010).

In trying to fuse the ideas of therapeutic, broad-spectrum, and active inoculation, there are several issues to consider, including the fact that in reality it is often not feasible to try to control people’s information diet or gauge their stance on every particular issue for which misinformation exists. Indeed, we do not always know people’s “infection” status across issues, but we do know that in the best-case scenario therapeutic inoculation can still confer immunity regardless. However, asking people to read a 600-word essay is not going to be a scalable intervention, so to bring inoculation into the digital era, we decided to create a real-world intervention that would make inoculation theory suitable for deployment “in the wild” and improve the ecological validity of misinformation research. Accordingly, we created the first fake news game: *Bad News* (Roozenbeek & van der Linden, 2018, 2019).

4.1 Bad news

The *Bad News* game (Roozenbeek & van der Linden, 2018, 2019) is an interactive social media simulation (the game simulates *Twitter*) which allows players to step into the shoes of a fake news tycoon and over the course of several levels produce their own misleading news media content (see Fig. 2).

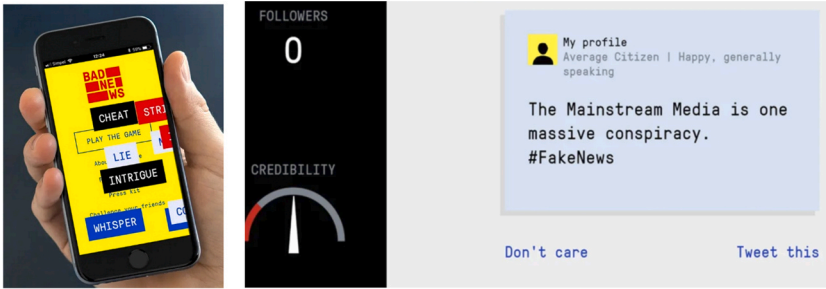


Fig. 2 Screenshot of landing page and gameplay (www.getbadnews.com).

As such, *Bad News* is a ~15-minute “choose your own adventure” (serious) game where players are presented with headlines and multimedia content that they can react to and share with other simulated users.

The goal of the intervention is to amass as many followers as possible without losing credibility. Players achieve this by making use of controlled weakened doses of the techniques used in the production of misinformation. The game features six techniques, including (1) the creation of *polarizing* headlines to exacerbate existing societal tensions and drive groups apart, (2) floating *conspiracy theories* by casting doubt on the mainstream narrative, (3) using *negative emotions* to fearmonger and cause outrage, (4) *discrediting* fact-checkers and the mainstream media by using deflection and denial, (5) distorting public discourse by *trolling* people online and finally, (6) *impersonating* experts and politicians to dupe people into believing and spreading misinformation. These techniques are not an exhaustive list but based on extensive reviews of the literature, intelligence reports, and interviews with professional producers of misinformation (for detailed accounts, see [Roozenbeek & van der Linden, 2020, 2022, 2023; van der Linden, 2023a](#)). Once a player completes a level, they receive a badge and move on to the next level. The game ends when players have mastered all techniques. If players do not follow the path of a clever manipulator, they lose all credibility (people can always opt out and “die a hero” if they feel uncomfortable with the gaming experience). The game was designed in collaboration with a design studio (Gusmanson) and media literacy organization (DROG/TILT) so we had to negotiate a balance between having a fun educational intervention but also a tool that can be used for scientific research.

The theoretical underpinning of the intervention is based on the process of *active inoculation* or letting players generate their own resistance by engaging with weakened doses of “attacking” material in a simulated environment to

help motivate immunity against a wide range of misinformation. Through a perspective-taking exercise players step into the shoes of an “evil manipulator”, which is meant to immediately elicit the threat component of inoculation. Much has been written about the role of threat in inoculation, which in contemporary inoculation scholarship is not meant to scare people but rather to motivate them to resist manipulation (Banas & Richards, 2017). In the game, players are not only forewarned about the tactics that manipulators use but they also receive weakened dose examples of what online manipulation looks like, often using humor, sarcasm, and ridicule. Just as the immune system requires exposure to many copies of a potential invading pathogen to mount an effective immune response, people need to be exposed to a wide range of examples of how common manipulation techniques are applied in order to spot and neutralize them (van der Linden, 2023a). From an ethics perspective, the content is all fictional to avoid the possibility of accidentally spreading misinformation and the game does not teach people how to write fake news nor how to make money from doing so (to eliminate any such incentives). Instead, it lets players experiment with weakened doses of propaganda techniques that are used to fool people. In McGuire’s terms: to trigger but not overwhelm the immune system (McGuire, 1964). Research on resistance to persuasion shows that people are more likely to generate resistance when (1) they perceive manipulative intent and (2) recognize their own vulnerability in the process (Sagarin et al., 2002) so the game scenarios and experience heavily leverage a focus on duping people and how to detect manipulation.

Because the game covers many topics and issues (from climate change and GMOs to gun control and missing airliners), it does not make sense to assess people’s baseline attitudes toward specific issues as was done in traditional inoculation research. In fact, the focus of the game is not on the issues themselves but about the larger manipulation techniques at-play, for which it is nearly impossible to measure prior exposure (e.g., people might not even be aware that they have been exposed). We *can*, however, measure people’s baseline susceptibility to such techniques. In particular, we hypothesized that familiarizing people with micro-doses of the techniques used in the production of misinformation should generate broad-scale immunity against a whole range of “full-dose” misinformation that makes use of these techniques regardless people’s political leaning or pre-existing attitudes. In the following section, I review some key empirical studies (along with their limitations) which have tried to evaluate the efficacy of this real-world active and generalized inoculation intervention.

4.2 Good news about bad news

Part of the novelty of the *Bad News* game is the ability to survey people in a simulated social media setting which helps enhance the ecological validity of the task given that people are typically asked to rate social media posts or headlines in a setting that is fairly detached from people's everyday social media experience. However, because surveys significantly slow down the engine and gaming experience, we initially were constrained by the fact we could only administer a limited number of test items. When the game was launched publicly it went viral in the mainstream media (van der Linden, 2023a) which allowed us to gather fairly "big data" as part of a naturalistic field experiment. To evaluate the degree to which people had been inoculated by the game experience, we included a pre and post-test in the initial studies to assess both people's baseline susceptibility toward misinformation and whether and how much people had improved post-gameplay.

Rather than administering fake headlines that had been featured and fact-checked in the media, which is often the predominant approach in the literature (see Pennycook, 2023; Pennycook & Rand, 2019), we decided to start off by creating our own headlines for two important reasons. First, there are major memory and source confounds with existing headlines so that people might have seen or heard about "real" fake news before and thus simply know whether it is credible or not regardless of the intervention (Roozenbeek & van der Linden, 2019, 2020). The second reason is that we wanted to exert experimental control and isolate the relevant manipulation techniques in each respective headline (and not others). Having said this, the test items were modeled after real-life examples to strike a good balance between experimental control and ecological validity. An example item for the conspiracy technique would be: "*#The Bitcoin exchange rate is being manipulated by a small group of rich bankers. #InvestigateNow*". Notably, the test items were different from the weakened doses people were exposed to in the game (i.e., refutational-different inoculation). Fig. 3 provides an example of the test environment. Consistent with the definition provided in Section 1.1 and in contrast to much existing research (van der Linden, 2022), we did not ask people whether they thought a headline or post is simply true or false but rather to calibrate their judgments on a scale of 1–7 based on how reliable people find the headline. We also included a couple of "credible" news items which did not contain any manipulation, such as a headline about the official #Brexit date. People rated the same items twice once before and once after playing the game.



Fig. 3 Misinformation test item example in the *Bad News* game. *Note:* The top panel illustrates how a technique [*impersonation*] is used in the game, and the bottom panel shows how the same technique is used in a different example on which participants were evaluated before and after playing. *Adapted with permission from Roozenbeek et al. (2019).*

We opted for a simple within-subject design because it was a live field experiment that allowed us to gather about 15,000 paired responses in the months following the public launch (these analyses were exploratory based on a unique media opportunity). The results confirmed that when exposed to the “full dose”, people became relatively more immune to a wide range of misinformation they hadn’t seen before, as indicated by a significant decrease in the average reliability rating of manipulative items ($d = -0.52, p < 0.001$) with some variation when looking at the specific techniques, ranging from $d = -0.16$ (polarization) to $d = -0.36$ (conspiracy). Although statistically significant due to the large sample size, there was no practically meaningful impact on the credible items ($d = 0.03$). Although conservatives performed worse at baseline consistent with much evidence in the literature (e.g., [Garrett & Bond, 2021](#)), there was no difference in the inoculation effect between liberals and conservatives ($d = 0.02$), likely due to the politically balanced nature of the game (players can craft scenarios that are

congenial to spreading misinformation about either liberals or conservatives).³ We did find that those who performed worse on the pre-test were much more likely to improve at post-test ($d = -0.89$) consistent with the notion that those who need the vaccine the most also benefit the most (Roozenbeek & van der Linden, 2019).

Of course, there are various methodological critiques to consider when interpreting these results, including the small number of test items (which could indicate item-effects), the lack of randomization (which could undermine causal inference), and the fact that the online sample was entirely opt-in (skewed toward male, liberal, and higher educated individuals). To remedy these shortcomings, we conducted a randomized experiment with a Prolific sample assigning participants to either the *Bad News* game or a gamified control group who played Tetris (Basol et al., 2021). We also implemented various quality control measures such as the fact that participants required a password to complete the study which would only become available if they finished the game in its entirety. We also added more misinformation items so that reliability of the items could be assessed, three for each of the six techniques for a total of 18 headlines ($\alpha = 0.84$). We also added a new hypothesis that the game might help boost people's confidence in their ability to resist manipulation (Tormala & Petty, 2004). After all, if people are not very confident in their abilities, they might be easily persuaded. We confirmed that, on average, inoculated participants found misinformation less reliable ($d = -0.61$, $p < 0.01$) with effect-sizes ranging from $d = 0.14$ (polarization) to $d = 0.58$ (discrediting). We again found no significant interaction between condition and ideology. Consistent with our hypothesis, players also became more confident in their reliability ratings ($d = 0.52$) compared to the control group (Fig. 4). One could argue that there is a risk of making people *overconfident*, but notably this boost in confidence was only observed for those who updated their prior ratings broadly in the right direction (i.e., less reliable).

We then conducted three more pre-registered confirmatory replications with the aim of documenting the decay of the acquired immunity over time (Maertens et al., 2020). The experimental setup remained consistent: we either randomly assigned people to *Bad News* or a gamified control group and participants completed the 18-item rating task both before and

³In later research using behavioral in-game data, we found rather mixed evidence for the hypothesis that players mostly craft scenarios that are congenial to their self-reported ideology (see Harrop et al., 2023).

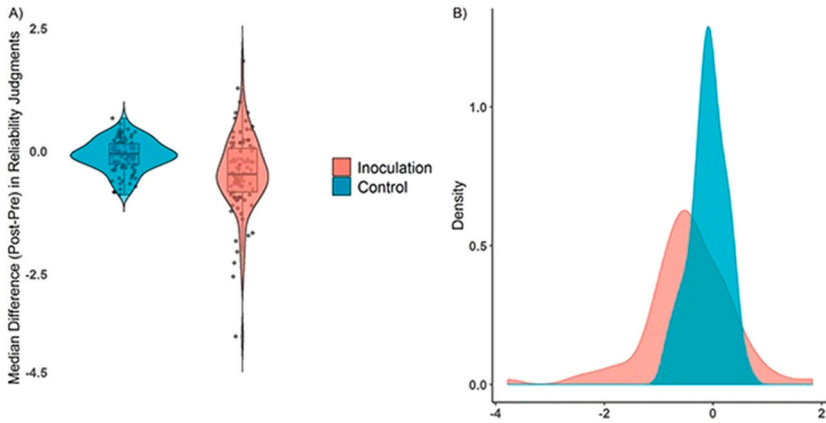


Fig. 4 Median change scores (post-pre) in reliability judgments across treatment conditions with jitter (Panel A) and density plots of the data distributions (Panel B). *Bad News RCT results, reproduced with permission from Basol et al. (2020).*

after the intervention in both groups. The novel twist in this set of experiments is that we “attacked” people with the misinformation item rating task at regular intervals (1 week, 5 weeks, and 3 months later) in the first experiment and only at the very end (13 weeks) in the second experiment. What we found was quite interesting, namely that we once again replicated the inoculation effect ($d = -1.0$) but with no significant decay over time. Yet, in the second experiment ($d = -0.69$), we did find significant decay of the inoculation effect of nearly 65% at 2 months ($d = -0.35$, Fig. 5). The only difference between the two experiments is the intermediate testing, which led us to hypothesize and later confirm (see Section 5) that misinformation quizzes (as well as re-exposure to the treatment) can actually serve to “boost” people’s immunity over time by strengthening two core processes that underlie inoculation: (1) renewed motivation to defend oneself from fake news and (2) enhanced memory of the lessons on how to spot it (Maertens, 2023; Maertens et al., 2023; Maertens et al., 2023).⁴

⁴ I should note that inoculation is about more than just converting the preemptive refutations to long-term memory as the refutational-different work shows that people can resist misinformation that was not explicitly mentioned in the prebunk, evidencing “true learning” (Compton, van der Linden et al., 2021).

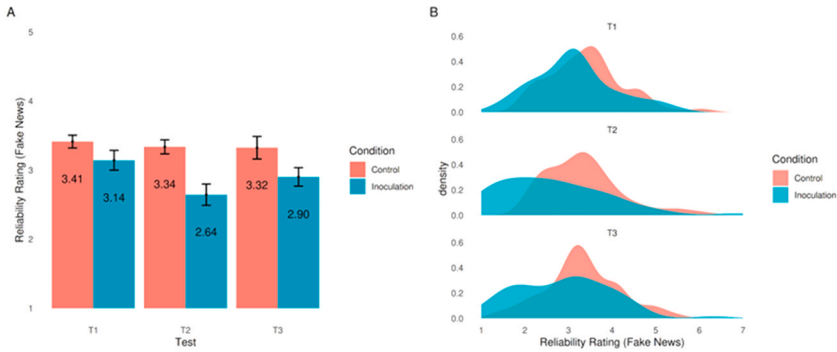


Fig. 5 Decay of the inoculation effect. *Note:* Reliability ratings of the misinformation items (averaged), separated by time and condition. Panel A displays the average reliability ratings and Panel B the density plots. T1 = pre-test; T2 = immediate post-test (0 months); T3 = post-test (2 months). $N = 110$. Error bars represent 95% confidence intervals. *Reproduced with permission from Maertens et al. (2020).*

4.3 Item and testing effects

One issue that has come up is whether our results could be due to item or testing effects given that people answer the same questions multiple times. In Maertens et al. (2020, Experiment 3), we used different items at follow-up to help exclude item-memorization effects and indeed found that the inoculation results were exactly the same. However, to more rigorously explore these questions, we ran a pre-registered Solomon three-group design (Roozenbeek, Maertens et al., 2020) across two different studies (total $N = 2159$). In this type of design, participants are allocated to one of three conditions, Group 1 is a traditional within-subject design with a pre-test and post-test, Group 2 is administered the pre-test and post-test without an intervention (the control group), and Group 3 receives the intervention and post-test but no pre-test. This setup allows us to isolate the unique impact of the pre-test, the intervention, and the interaction between the pre-test and the intervention. As is visible from Fig. 6, there was a tiny yet non-significant difference in the control group (i.e., the pre-test effect is statistically equivalent to zero). A significant inoculation effect is observed in the pre-post-intervention condition (Group 1, $d = -0.36$). When we subtract the small pre-test effect from the post-test mean of Group 1 (the control group), we find that it is statistically equivalent to the post-test only group (Group 3), confirming no significant interaction between pre-test and the intervention. In fact, the inoculation effect remained highly significant when both the pre-test and interaction effects

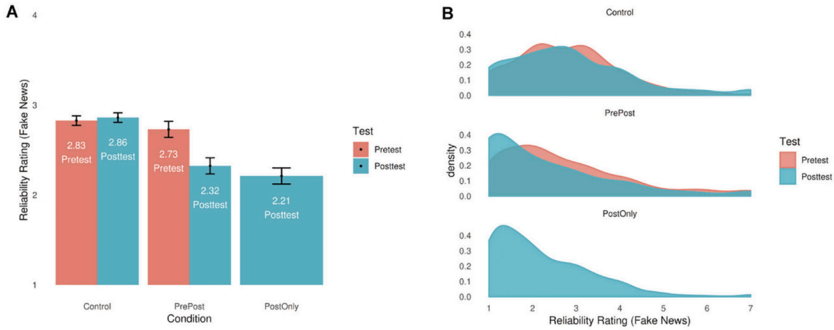


Fig. 6 Bar chart (A) and density plots (B) of fake news reliability ratings. $N = 1679$. Error bars represent 95% confidence intervals. *Testing Effects in the Bad News Game*, reproduced with permission from [Roozenbeek et al. \(2020\)](#).

were subtracted from the post-test only group, confirming that there are no testing effects in our design ([Roozenbeek, Maertens et al., 2020](#)).

The item effects picture proved more complicated. To look at item effects we assigned participants to receive different misinformation item sets (Set A or Set B) during both the pre-and post-test. If participants got Set A in the pre-test they would see Set B in the post-test and vice versa, so that Group 1 (A-B) saw Set A during the pre-test and B on the post-test and Group 2 (B-A) saw Set B on the pre-test and Set A on the post-test. When crossing the items between groups, we found a significant effect ($d = -0.40$) for Set A (pre-test Group 1 to post-test Group 2) but not for Set B ($d = -0.07$). When looking purely within groups, we do find significant effects (pre-post) for both Group 1 ($d = -0.19$) and Group 2 ($d = -0.29$). Yet, it is evident that overall, there is clear variation in the size of the inoculation effect depending on which item sets are used ([Roozenbeek, Maertens et al., 2020](#)).

When extrapolating these issues to the real-world, one critique that often comes up is the fact we create the item sets ourselves. Would the inoculation effect still hold if we use misinformation headlines that occur “in the wild” or even misinformation which does not necessarily contain the manipulation techniques people were inoculated against? In other words, can the *Bad News* game confer “cross-protection” against related misinformation that was not specifically addressed in the intervention? We explore these questions in two studies ([Roozenbeek, Traberg, & van der Linden, 2022](#)).

In the first experiment, we leverage the *Bad News* environment for a standard within-subject experiment but this time with examples from the real-world which explicitly use one of the six manipulation techniques. For example, for conspiracy we used a real example which read: “*Exposing the shadow elite controlling the world*” (from *Humans Are Free* – a conspiracy outlet) and for polarization: “*Professor calls right-wing reporting a form of violence*” (*InfoWars*). We found that, on average, the intervention had an effect similar to the fictional items in reducing people’s overall reliability ratings of misinformation ($d = -0.37$, range $d = -0.12$ to $d = -0.27$ per technique)⁵ suggesting that at least to some extent the items we use generalize and are representative of real-world misinformation.

In the second experiment, we conducted a semi-adversarial collaboration with Gordon Pennycook and David Rand. For this experiment, we return to the item-effect design from [Roozenbeek, Maertens et al. \(2020\)](#) and administer one of two item sets (Set A or Set B). We also changed the nature of the stimuli by selecting items not on basis of manipulateness but rather based on whether they are true or false. These items came from [Pennycook and Rand’s \(2019\)](#) database and are widely used and pertain mostly to U.S. political news. We also changed the wording of our dependent variable from “reliability” to “accuracy” to fit Pennycook and Rand’s preferred question wording and finally, we included a measure of veracity discernment as both item sets were balanced with 4 real and 4 fake headlines (e.g., *Denzel Washington: With Trump We Avoided War With Russia and Orwellian Police State*). Overall, when crossing the items, we found a significant inoculation effect for Set A ($d = -0.20$) but not Set B ($d = -0.03$). Across both sets within groups, we find a significant inoculation effect ($d = -0.10$) which is driven mostly by Group A-B ($d = -0.24$) compared to Group B-A ($d = 0.05$). There were no significant effects of the intervention on real news (some items increased, others decreased) with significant discernment ($d = 0.08$, $p = 0.002$).

Overall, what we conclude from this is that the *Bad News* inoculation game does protect against real-world misinformation and renders some cross-protection against false headlines that may not even contain the manipulation techniques that people were inoculated against. However, the effect-sizes are clearly small in this case and decrease the more the headlines differ from the

⁵We have replicated the equivalence between fictional and real-world misinformation items head-to-head in other games too (e.g., see [Neylan et al., 2023](#); [Roozenbeek & van der Linden, 2020](#)).

inoculation training in the intervention. Moreover, it appears that significant item-effects are at play given the range of effect-sizes across different kind of news stimuli (Roozenbeek et al., 2022).

4.4 GoViral, harmony square, and other inoculation games

Following the evaluations of *Bad News*, we were approached by the UK Government during the pandemic to create a special version of the game to help inoculate the public against pandemic misinformation. This allowed us to investigate the real-world generalizability and scalability of our interventions in new and interesting ways. We created a new, much shorter 5-minute game called *GoViral!* (Fig. 7) which had three levels based on the most frequent techniques used to spread misinformation about Covid-19 (fearmongering, fake experts, and conspiracy theories). In the game, players join a chatgroup called “not co-fraid” and start experimenting with the idea that kiwis secretly cure the coronavirus and fake their credibility by posing as a doctor. The game became part of the World Health Organization (WHO)’s official “*Stop the Spread*” and the United Nation (UN)’s “*verified*” campaigns and reached over 200 million impressions on social media (GCSI, 2022). To evaluate the intervention empirically, we employed both the typical “in the wild” pre-post design for the field study during the campaign period as well as three pre-registered parallel randomized experiments in the UK, Germany, and France. At the time, the UN also released prebunking infographics on social media as part of a separate campaign, which allowed us to test the relative efficacy of “*active*” versus “*passive*” inoculation in a naturalistic setting.

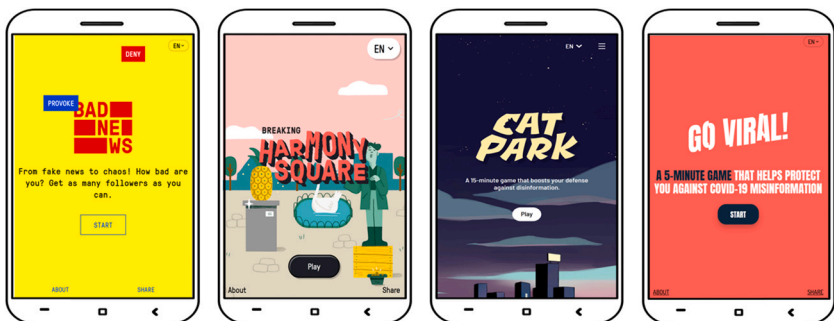


Fig. 7 Landing pages for the *Bad News* (<https://www.getbadnews.com/>), *Harmony Square* (<https://www.harmonysquare.game/>), *Cat Park* (<https://www.catpark.game/>) and *Go Viral!* (<https://www.goviralgame.com/>) games. Adapted with permission from Roozenbeek and van der Linden (2023).

During the first month of the release, 14,755 people participated in the online study, of which 2634 completed the pre-post test (about 18%, which is typical for our “live” interventions). We found that people rated COVID-19 misinformation as more manipulative post-gameplay ($d = 0.52$) and they also discerned better between manipulative and non-manipulative social media posts ($d = 0.36$). In the randomized experiments, we found that both the active (*GoViral*, $d = 0.56$) and passive (UN infographics, $d = 0.17$) inoculation campaigns were effective compared to control, but the game had a significantly greater effect-size than the infographics. These results replicated in each country as well. Interestingly, the UN campaign did not reduce intentions to share misinformation compared to control, but the *GoViral!* game did ($d = 0.15$).⁶ We also included a 1-week follow-up for the UK sample and found that although the effect of the passive inoculation was no longer significant, people who played the game still found misinformation less reliable one week later ($d = 0.27$). We also included various exploratory measures to better understand the relative potential benefits of active inoculation and found that people were more likely to want to share the game with others in their network than the infographics ($d = 0.19$) and experienced greater motivational threat (the motivation to want to defend against misinformation) in the *GoViral* condition compared to both the infographics and control conditions ($d = 0.15$) suggesting that active inoculation might motivate greater resistance to misinformation. Though I should add that our focus was on comparing real-world campaigns, there is of course the confound that the campaigns were very different in terms of their content, though more controlled experiments in other contexts also seem to favor active over passive inoculation (Green, McShane, & Swinbourne, 2022).

In 2020, we were working with the Department of Homeland Security and the State Department to help Americans recognize foreign disinformation techniques during elections, especially the 2020 U.S. Presidential election. To achieve this, we created another game with the U.S. Global Engagement Center and CISA (the Cybersecurity and Infrastructure Agency) called *Harmony Square* (Fig. 7), which is uniquely focused on exposing people to a weakened simulation of how foreign actors try to disrupt discourse during elections. These techniques were drawn from CISA’s pineapple pizza campaign, which more or less explains foreign influence operations through the

⁶ These effects were only significant when collapsed across countries, as it is typical to observe floor effects for sharing measures because people do not report to share much misinformation (hence it requires greater power to detect).

analogy of sowing chaos and discord about a seemingly innocuous issue such as whether or not to put pineapple on pizza. The five techniques include trolling people, using emotional and inflammatory language, false amplification (e.g., through bot armies), creating conspiracy theories, and polarizing audiences. In the game, *Harmony Square* is a peaceful fictional town where democracy is celebrated at the annual pineapple festival until you start meddling with a local election using weakened doses of key propaganda techniques, which are meant to trigger subsequently resistance.

Using an international sample ($N = 681$) and the same randomized mixed design used in all previously discussed studies, we found that after playing the game people found both fictional ($d = 0.51$) and real-world ($d = 0.54$) polarizing misinformation (e.g., “1.5 less MAGAbilly’s in the world. At least they died supporting their beloved 2nd amendment”) less reliable. They also became more confident in their assessments ($d = 0.30$) and significantly less willing to share such misinformation with others ($d = 0.28$).

This program of research inspired several other active inoculation games designed by other research groups, including the popular *Cranky Uncle* game (Cook et al., 2022), which inoculates people specifically against the rhetorical techniques used in climate denial, such as cherry-picking data and the use of fake experts. Another recent intervention, *Spot the Troll* (Lees et al., 2023) focuses on helping people spot signs of inauthentic accounts. In a nationally representative study, the authors found that the quiz (compared to a gamified control) significantly improved “troll discernment” (i.e., distinguishing authentic from inauthentic accounts), especially for those who were least discerning at baseline. The authors also included an observational field study as a subsample of the participants posted their score on Twitter. Using a three-week window, the authors noted an overall reduction in tweeting behavior among those who were in the intervention group, particularly in terms of retweeting (26% reduction in retweets per day). The authors speculate that the quiz might make people feel less confident in determining which accounts are authentic and as such lead people to exercise more general caution in deciding what and whether to retweet.

Yet, in these type of observational social media studies, it is often difficult to get a clear signal for low-quality information sharing, because baseline sharing is usually low and proxies for inauthentic accounts or polarizing language may not be granular enough. In addition, the problem of creating lists of users who share such content and getting the time window correct for measuring when such content is actively shared is

another challenge. To try to get around this problem, [McPhedran et al. \(2023\)](#) created a novel simulated social media setting where participants could share and engage with content (including ‘likes’ and various other emoticons). In a large national sample ($N = 2430$), they evaluated (1) an inoculation intervention (where participants were forewarned about the dangers of misinformation and given some cues on how to spot it across different domains), (2) a false tag intervention (broadly representative of how social media companies typically deal with misinformation) or a (3) baseline control. Findings showed that the inoculation was most effective and inoculated participants “liked”, “loved”, and “shared” misinformation posts significantly less than both the control and the false tag interventions (interestingly, there were some item effects with greater inoculation effects for health-based than political headlines).

4.5 Scaling inoculation

Games and quizzes are not the only vehicle or “virtual needle” available to deliver the inoculation to participants ([Compton et al., 2021](#)). Like text-based stimuli, videos are considered “passive inoculation” as participants do not actively create content or counterarguments themselves. Nonetheless, videos are easier to scale on social media than large articles of text. In a unique collaboration with Google, we studied misinformation techniques that are prevalent on video platforms such as YouTube. Rather than using headlines, political gurus often resort to misleading rhetorical techniques to incite extremism such as false dichotomies, scapegoating, fearmongering, incoherence, and ad hominem attacks ([Lewandowsky & Yesilada, 2021](#)). We created four short videos that followed an inoculation script closely: people were first forewarned that manipulators make use of these techniques followed by exposure to a weakened dose example along with an explanation of how to spot them. In five pre-registered experiments, [Roozenbeek et al. \(2022\)](#) tested whether the inoculation videos helped improve recognition of these techniques, people’s ability to discern between manipulative and non-manipulative posts, how trustworthy they deem such posts to be, the level of confidence in their ability to make these judgments, and finally their intention to share social media posts which contain these techniques.⁷ Participants were recruited via Prolific and

⁷ We ran a separate experiment randomizing the order of these outcome variables and determined there were no order effects (e.g., asking about sharing first or last did not impact people’s ratings).

randomly assigned to one of the videos (1 per experiment) or a control group (who watched an unrelated video). In terms of the stimuli, in each experiment people were shown ten social media posts, each of which was randomized to be either manipulative (using one of the relevant techniques) or neutral (matched in content but without using any manipulation). Different participants saw different specific stimuli but on average, all participants rated five manipulative and five neutral items. All source information was blocked out to avoid source confounds. An example item for false dichotomies would be “*We either need to improve our education system or deal with crime on the streets.*”

The results are displayed in [Fig. 8](#). The videos consistently improved people’s (1) ability to discern between manipulative and non-manipulative content ($d = 0.28$ to $d = 0.62$), (2) trustworthiness ratings ($d = 0.10$ to $d = 0.25$), (3) the level of confidence in their ratings ($d = 0.24$ to $d = 0.50$) with the exception of incoherence ($d = 0.04$), and (4) sharing discernment for three out of the five videos ($d = 0.10$ to $d = 0.22$)—keeping in mind floor effects for sharing. These effects were not moderated by political ideology or conspiracy mentality.

Other independent research teams who collaborated with Google to evaluate the efficacy of inoculation interventions obtained similar results. For example, [Piltch-Loeb et al. \(2022\)](#) found in a randomized study with unvaccinated individuals ($N = 1991$) that inoculation videos that focused on similar manipulative tactics significantly improved participants’ ability to detect the misinformation techniques, lowered willingness to share such misinformation, and increased their propensity to get the COVID-19 vaccine. In addition, [Pennycook et al. \(2023\)](#) replicated the findings from [Roozenbeek et al. \(2022\)](#) for the emotion videos but showed that in addition to techniques, true-false discernment can be boosted too by adding an accuracy nudge on top of the inoculation treatment.

Although these effects were consistent in laboratory settings, we know from other research that effect-sizes are often smaller in the field ([DellaVigna & Linos, 2022](#)). In the final and sixth experiment of [Roozenbeek et al. \(2022\)](#), we therefore conducted a field study in collaboration with Google and YouTube. We used two videos in a real advertising campaign on the YouTube platform and about 5 million U.S. users (consumers of political news) were exposed to the campaign as an ad while watching content on YouTube. About 1 million users watched the videos for at least 30 s or more. The campaign was randomized so that people either saw one of the videos or

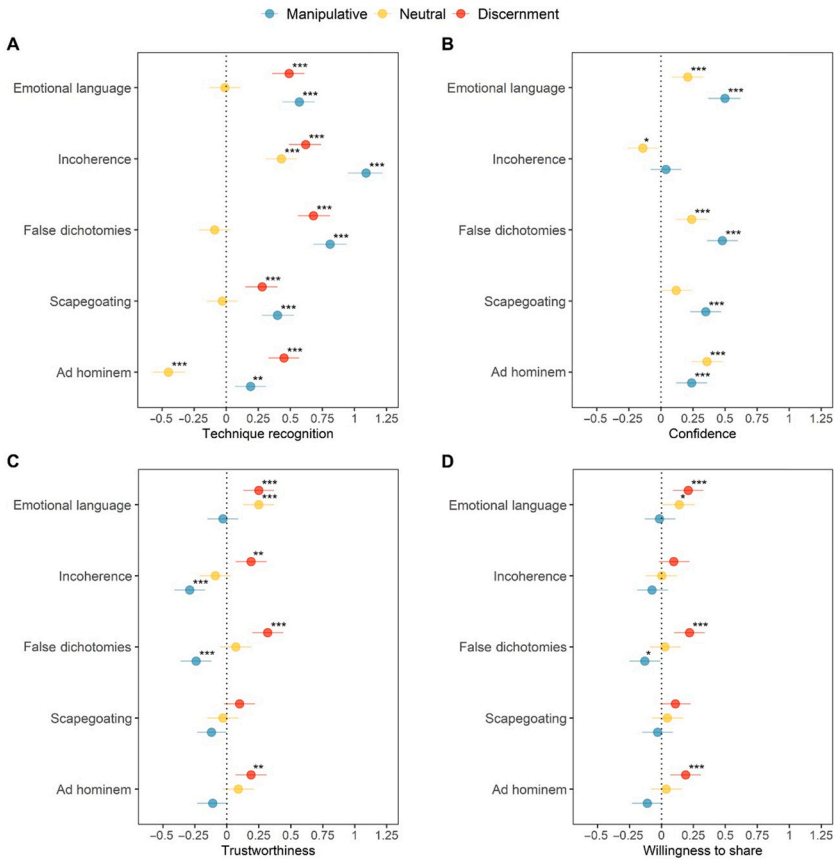


Fig. 8 Video inoculation Studies 1–5: Dot plots for effect sizes (Cohen's d). Note: Technique recognition (A), confidence (B), trustworthiness (C), and sharing (D), by condition and study, for manipulative social media content, neutral content, and discernment. Improved discernment indicates a higher ability to distinguish manipulative from neutral stimuli. Error bars show 95% confidence intervals for Cohen's d . *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Adapted with permission from Roozenbeek et al. (2022).

a control video. About 30% of the treatment group was then randomly selected for a survey within 24 h after exposure to the ad (the median time lapse between exposure to the inoculation video and our test was 18 h). The survey contained a fictional social media post and users answered a multiple-choice question to indicate which technique they thought was being used

(there was only one correct answer).⁸ We could only ask one question in the YouTube brand lift environment (the surveys that pop up after seeing an ad) but we made sure to use different items to avoid item effects (3 items per video). On the whole, in this more ecologically valid setting where users are distracted on social media, we increased technique recognition consistently between 5% and 10% relative to control (Cohen's $h = 0.09$). This is the largest field validation of inoculation theory on social media that we know of to date (Roozenbeek et al., 2022). Yet, many challenges of course remain to conducting such trials on social media (van der Linden, 2023), including obtaining behavioral data. For example, we might wonder if people who were exposed to the inoculations went on to watch less misinformation, but YouTube does not publish behavioral data and did not want to make viewing records for individuals available for scientific research. Yet we know from McPhedran et al. (2023) that at least in simulated environments, inoculated users share less misinformation with others though more randomized studies on social media are necessary (van der Linden, 2023b), especially because people often rely on social and source cues when judging whether to share content which can (partially) detract from the inoculation (Traberg, & van der Linden, 2022; Traberg et al., 2022).



5. Mechanisms: The memory-motivation model of inoculation theory

To combine our insights from text-based, game-based, and video-based inoculation studies, we decided to investigate a more comprehensive theory that can explain what mechanisms drive the effectiveness and longevity of inoculation against misinformation. Two conceptual measures are common in this literature corresponding to each of the two elements of inoculation: threat and counter-arguing. Scholars are particularly divided over the role of threat as a mechanism. Whereas it was initially conceived (but not measured) as an implicit feature of the weakened dose which McGuire (1961a) described as having “shock value” (p. 185), he later experimented with more explicit forewarnings (McGuire & Papageorgis, 1962). Compton and Ivanov (2012) found that both the explicit forewarning

⁸ The multiple-choice format is a well-known limitation of the brand lift environment in YouTube which is not setup for scientific or psychometric testing and is usually reserved for ad recall and brand recognition polls.

and the weakened dose can generate perceived attitudinal threat. Some scholars have even argued that “threat is the most distinguishing feature of inoculation” (Pfau, 1997, p. 137) though others have wondered how exactly it fits with the biological analogy (see discussion in Compton, 2021).

Somewhat surprisingly, meta-analyses, which have found strong evidence that inoculation confers resistance to persuasion ($d = 0.43$), have not found much evidence for the moderating role of threat (Banas & Rains, 2010). One interesting potential explanation has been that existing measures of attitudinal threat are not tapping into the right construct (Banas & Richards, 2017). For example, Basol et al. (2021) found that traditional measures of “apprehensive threat” measured with bipolar adjective scales (such as threatening–nonthreatening) appear less relevant than “*motivational threat*” or measures that motivate an “immune” response and thereby tap into people’s desire to want to defend themselves from impending persuasive attacks (e.g., “I feel motivated to resist misinformation”).

The second measure most commonly used to assess the efficacy of the weakened dose and subsequent refutations is the extent to which people self-report to want to counterargue against the (misinformation) attack (McGuire & Papageorgis, 1961). Yet, surprisingly little research has examined the extent to which people remember the refutations. One study did examine the role of associative memory structures where the authors found that prebunks (or refutational pre-emptions) can modify the structure of associative memory networks by increasing the number of nodes (concepts) in the network and strengthening the linkages between nodes leading to greater resistance to persuasion (Pfau et al., 2005).

Yet, the key overarching question is why inoculation loses its effectiveness over time? Either people gradually lose motivation to defend themselves from future persuasive attacks or they lose the ability to defend themselves by forgetting the inoculation material (or perhaps both). In a recent study (Maertens et al., *in revision*; Maertens, 2023), we revisit both mechanisms across five pre-registered longitudinal experiments (total $N = 11,759$) where we independently measure and manipulate both mechanisms over 30 days (with three measurement points) across different types of inoculation treatments: text (a replication of the climate change experiment), games (*Bad News*), and videos (from the Google study). In most of these experiments, participants receive a pre-test and are subsequently randomly assigned to either the control group, an inoculation treatment, or a booster group (which received a short booster of the initial inoculation treatment). We test three theoretical accounts: one where the

inoculation effect is sustained solely via threat (Fig. 9, Model A), which we refer to as the "threat-motivation" model; one where the effect is mostly attributable to memory and learning, i.e. the "memory model" (Fig. 9, Model B); and finally, an integrated model that combines both mechanisms and postulates that although memory is the leading mechanism, people need to be motivated to learn and remember the inoculation material (Fig. 9, Model C).

We included a large number of items representing each measurement construct to avoid concerns about measurement. For example, we assessed both objective memory of the treatment material (e.g., correct recall of the relevant micro-dose or prebunk via a multiple choice question), self-reported memory of the material, associative concept mapping (written responses), motivational threat (e.g., "thinking about emotional manipulation motivates me to resist misinformation"), apprehensive threat (e.g., "I feel threatened"), as well as other potentially relevant factors such as issue involvement.

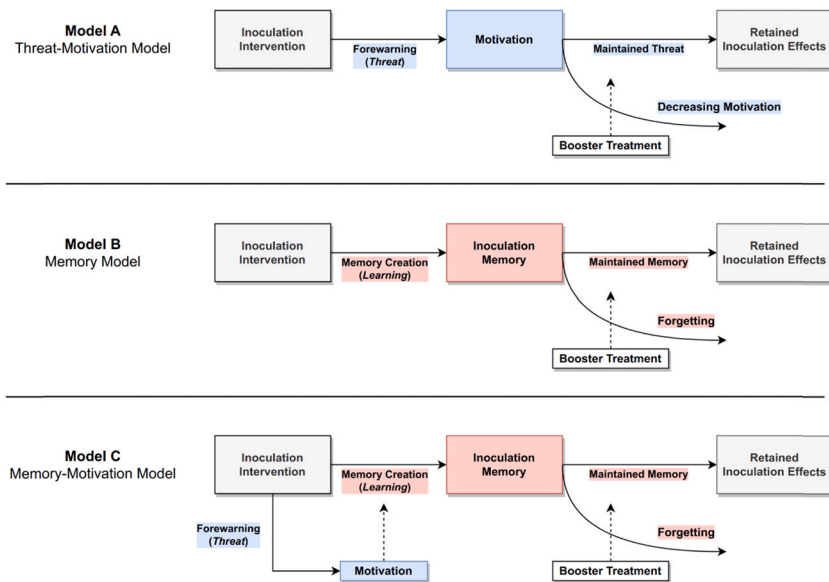


Fig. 9 Memory-motivation model of inoculation theory. *Note:* Model A assumes the main mechanism of inoculation is threat and that people lose the motivation to defend themselves from the threat of misinformation over time. Model B assumes that the main explanation is cognitive or loss of memory of the inoculating material. Model C integrates both accounts arguing that while memory is mainly responsible for long-term retention of resistance to misinformation, motivation elicited by threat helps boost people’s memory. *Adapted with permission from Maertens et al. (2023).*

Across all experiments we found memory to be the single most important predictor of decay of the inoculation effect and boosting memory (but not threat) was able to help maintain longevity of the inoculation effect compared to the non-boosted conditions. In general, the decay function of inoculation interventions has remained elusive. For example, contrary to McGuire's initial assumption of a curvilinear relationship where people need time to build up cognitive resistance after the inoculation (and before the attack), most studies find that the inoculation effect decays over the course of a couple of weeks (Banas & Rains, 2010; Banas & Miller, 2013; Maertens et al., 2021). In our studies, we find strong evidence that the inoculation effect behaves much like an exponential forgetting curve (Murre & Dros, 2015) with significant decay over the course of a month unless boosters are implemented. Interestingly, these boosters can take many forms, including repetition of the material, a quiz, or even just another post-test, yet a booster solely focused on refreshing threat did not work.

Overall, those who displayed better memory of the material also showed increased resistance. Dominance analysis with all variables indicated that memory was consistently the variable that explained the most variance in the inoculation effect over time (41%–82% compared to 2%–27% for motivational threat, the 2nd best predictor). However, SEM models testing each of the three theoretical accounts did find support for the integrative model where motivational threat predicts increased memory both directly and indirectly (see Fig. 10). Yet, memory remained a stronger predictor with the memory booster being important for longevity of the inoculation effect. Somewhat unexpectedly, only the video interventions successfully manipulated threat, which could mean we did not sufficiently target threat (but this would be rather surprising as even the booster video that was specifically aimed at generating threat had no direct effect on motivation). We therefore leave open the question of finding better ways to motivate people to defend themselves from fake news, but we note that our gamified interventions, for example, already make people aware of their own vulnerability. In our current models, it seems that threat plays a more indirect role by motivating people to engage and remember the content from the inoculation, which then subsequently predicts resistance and longevity. Importantly, the fact that research finds that people show resistance to attacks or arguments not raised in the initial inoculation implies that people not only retrieve information from long-term memory but also engage in learning (Compton et al., 2021; Roozenbeek et al., 2022).

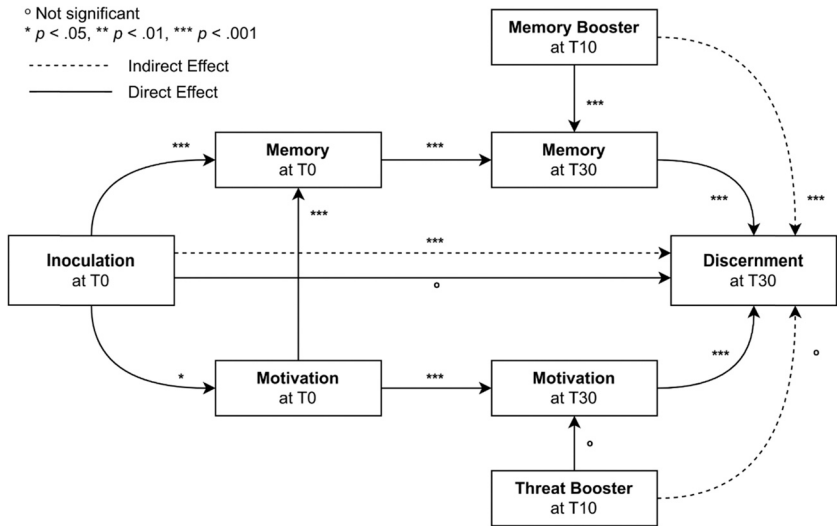


Fig. 10 Structural equation model (SEM) for the integrated account (Model C). *Note:* T0 = pre-test, T1 = post-test at 10 days, and T30 = post-test at 30 days ($N = 2220$). This model is most representative because it separates out the effect of the memory and threat boosters. Only memory “boosters” enhance people’s discernment of misinformation at T30 (30 days) while both memory and motivation processes predict discernment (both directly and indirectly).



6. Limitations and future directions

A recent meta-analysis (Lu et al., 2023) of inoculation theory in the domain of misinformation found strong support for the use of inoculation treatments in reducing endorsement of misinformation ($d = -0.36$, 95% CI $[-0.23, -0.50]$, $p < 0.001$) and improving veracity discernment ($d = 0.20$, 95% CI $[0.13, 0.28]$, $p < 0.001$). Still, many open questions about inoculation research in the context of misinformation remain, including the lack of cross-cultural research, conceptual confusion around the term “pre-bunking”, potentially undesirable side effects of the inoculation treatment, as well as scope for extending the analogy from (individual) cognitive resistance to (social) herd immunity.

6.1 Cross-cultural research

The overwhelming majority of inoculation and misinformation research (Badrinathan & Chauchard, 2023; Blair et al., 2023) has been conducted in

predominantly Western, Educated, Industrialized, and Rich Democracies (WEIRD). In the context of misinformation this seems especially prudent given that citizens in other parts of the world likely have very different relationships and histories with (state-controlled) media, propaganda, and censorship suggesting there could be important cross-cultural differences in the effectiveness of inoculation interventions. Through a collaboration with the UK Foreign Office, we were able to translate and adapt the *Bad News* game in many different languages around the world with the help of local media literacy organizations. In one study, using the standard in-game survey, we found the game to be effective in Greece, Germany, Sweden, and Poland with slightly greater effectiveness in Germany as compared to Greece and Poland, though we were not able to explain the heterogeneity (Roozenbeek, van der Linden, & Nygren, 2020).

Some of the consequences of viral misinformation have been particularly damaging in the Global South, including mob lynchings in India following the spread of false rumors on WhatsApp (Goel, Raj, & Ravichandran, 2018). A team of scholars recently translated and replicated the *Bad News* game in India using local headlines, finding very similar effect-sizes for most of the techniques as the original study (Roozenbeek & van der Linden, 2019) as well as a significant improvement in participants' ability to discern credible from manipulative news (Iyengar, Gupta, & Priya, 2022). As part of a research collaboration with WhatsApp we created a new game called *Join This Group* which is particularly focused on inoculating users against the viral spread of false rumors and the escalation of violence on direct messaging platforms. Although the game was effective when tested on a UK sample (Basol, 2022), we did not find any effects when we tested the game in rural India, neither in terms of detecting manipulative nor credible social media posts (Harjani, Basol, Roozenbeek, & van der Linden, 2023). One major difference between Iyengar et al. (2022) and our study is that the former used an educated urban sample whereas we focused on a rural sample with low literacy. We collaborated with the Digital Empowerment Foundation (DEF), a local media literacy group in India, who recruited 1283 individuals into the trial from 8 North Indian states. They went door-to-door with iPads and iPhones to administer the intervention and survey.

Significant difficulties occurred during the field work, including with the translation of the game, which could have resulted in loss of humor and understanding. In addition, we were told that the "threat" element of the game (elicited by stepping into the shoes of an evil propagandist) did not fit

the media literacy program of DEF as there were concerns that it could cause confusion, especially given the sensitive political climate in India. Accordingly, we had to adapt the intervention, and some scholars might wonder if the intervention still counts as “inoculation” without a clear “threat” component (Compton, 2021). Media literacy levels are also much lower in rural India (Badrinathan, 2021) and so participants may not have understood the purpose of the game, especially given differences in how people use social media in India versus Western countries (Banaji et al., 2019). Overall, we took these results as a valuable lesson about the importance of cultural adaptation given potentially large cross-cultural differences in people’s understanding of (social) media and misinformation.

Other efforts include Google Jigsaw, who replicated our YouTube video campaign study “in the wild” in the context of disinformation about Ukrainian refugees in Europe, reaching a majority of all social media users in Poland (50%–62%), Czechia (68%–80%), and Slovakia (55%–69%). The improvement in manipulation discernment revealed some cross-cultural heterogeneity, ranging from 1.9% to 8.1% (Jigsaw, 2023). The largest effects were observed in Poland with little meaningful effects in Slovakia. Some interesting potential moderators surfaced such as whether the videos were simply dubbed (as was the case for Slovakia) or culturally tailored (which was the case for Poland). A more general moderator seemed to be whether the algorithm optimized for viewing time (users watching the whole video) or reaching the largest possible audience, as effect-sizes were typically greater when the algorithm optimized for viewing time (Jigsaw, 2023). A second campaign in Germany, which reached over 40 million views, proved highly consistent and in line with the results of the original study (Jigsaw, 2023b).

In other work, we have tested the potential for inoculation in related domains such as preventing violent extremism. The playbook of violent extremism breaks down into very clear techniques against which people can be inoculated, including identifying vulnerable targets, isolating them from their friends and family, gaining their trust, and activating them to do something bad on behalf of a greater cause (Doosje et al., 2016). We created an intervention (“*Radicalize*”) in collaboration with Nudge Lebanon where players step into the shoes of an extremist organization and try to recruit people into a cause called the *Anti-Ice Freedom Front* (AIFF) whose mission is to melt the global ice caps. Players make use of weakened doses of the key strategies employed to radicalize people into extremism (Saleh et al., 2021). We initially tested the game on a UK sample with vignettes simulated via WhatsApp group messages and found that players

significantly improved in their ability to detect these techniques (Saleh et al., 2021). One critique that surfaced during peer-review was that we had not shown this intervention to be effective on actual at-risk youth. To remedy this limitation, we recently conducted a field study in post-conflict regions in Iraq that were formerly occupied by ISIS. The game was translated by native Iraqi Arabic speakers and we changed the name of the game to MINDFORT (“fortifying the mind”) to avoid any translational confusion with the term “radicalize”. We collaborated with Nudge Lebanon and a local organization (Spirit of Soccer) who organize workshops for vulnerable youth. During these workshops we were able to recruit 191 individuals representative of the target group (mostly men in their twenties) to participate in the study. Results replicated the main outcome variable of interest: identifying the manipulation techniques (Saleh et al., 2023), although the effect-size was 50% lower than in the original study ($d = 0.31$ vs. $d = 0.71$) and we did not replicate significant effects on secondary outcomes such as identifying vulnerable profiles, possibly because of low statistical power but also because local understanding of at-risk characteristics (e.g., unemployment, marginalization) may be normalized rather than seen as unusual as it would be in many Western countries (Saleh et al., 2023).

More research is emerging, such as the efficacy of active inoculation games in China where *Bad News* is banned, but local scholars have developed games inspired by *Bad News* via *WeChat* which have shown to (durably) increase discernment around (COVID-19) misinformation during the pandemic (Lu et al., 2023; Ma et al., 2023; but see Rędzio et al., 2023 for different results in Poland). Overall, it is clear that collaboration with local experts and organizations is absolutely key to understanding cross-cultural variation in the effectiveness of inoculation interventions, which should be a priority for future research.

6.2 Prebunking versus inoculation

The term “prebunking” is now increasingly used to describe a suite of anti-misinformation interventions, causing some confusion about its meaning across different kinds of studies (Roozenbeek et al., 2023; Traber et al., 2022). When we originally started using the term “prebunking”, we were referring to the refutational pre-emption phase of the inoculation process to distinguish it more clearly from debunking (van der Linden, Leiserowitz et al., 2017). Another popular use of the term “prebunking” is to refer to any type of intervention that simply comes *before* rather than after exposure to misinformation, including accuracy nudges and “think before you post” interventions

(Pennycook, 2023; Pennycook et al., 2020). Sometimes researchers even use the term “prebunk” or “inoculation” for a fact-check intervention where the only difference is the timing of a warning label (e.g., Brashier et al., 2021). This is an extremely watered-down use of the concept and not commonly understood as inoculation (Traberg et al., 2022). Other interventions are using the term “prebunking” to denote preemptive strikes that leverage other theories than inoculation, such as engaging in counterfactual thinking before exposure to a fake news post (Bertolotti & Catellani, 2023). Perhaps a good strategy moving forward is to differentiate prebunking from inoculation theory specifically although this does not necessarily resolve the use of “prebunking” as a core mechanism of the inoculation process. It is also apparent that the term “prebunking” is often preferred for political reasons by technology companies and governments to avoid eliciting negative associations with the vaccine analogy. So, while it is perhaps unavoidable that these terms will be used interchangeably, making conceptual distinctions clear will help scholars better understand, apply, and interpret the use of prebunking and inoculation theory.

6.3 Side effects: Skepticism, real news, and discernment

One important area of investigation is whether inoculation interventions produce any (un)desirable “side effects” akin to how some people might experience (mild) negative reactions from a vaccine (Compton et al., 2021; van der Linden, 2023a). As in medicine, an intervention might achieve its intended aim (e.g., boost immunity) but there are various potential side effects (e.g., rash, fever) that need to be considered before deciding whether the overall benefit exceeds any potential harm for a group or individual.

For example, one early concern about the *Bad News* game was whether letting people step into the shoes of a manipulator might inadvertently incentivize people to start spreading fake news themselves. Although we have not explicitly asked people about their intentions, we were able—using a case-control design—to scrape and analyze hundreds of comments on Reddit and trace which commenters recently completed the game (by posting their score) and found that the content of the comments was mostly geared toward applying and furthering the inoculation analogy rather than expressing any nefarious intentions (see discussion in van der Linden, 2023a).⁹ However, future research may want to examine this more directly, for example, by asking people about their intentions, tracking what websites people visit or

⁹Subgroups of more extreme populations might have politically motivated rather than accuracy-oriented goals (though nefarious actors tend to already know about these manipulation techniques).

what they post on social media after having completed an inoculation intervention. This would be especially fruitful in light of the lack of behavioral (field) data from inoculation interventions.

Another question surrounds the role of “confidence” and whether inoculation interventions could perhaps make people *overconfident* in their abilities, which would be concerning because overconfidence is linked to the sharing of fake news (Lyons et al., 2021). There seems to be little evidence for this at present. In one study, for example, confidence only increased among participants who had correctly downgraded their reliability ratings of misinformation post-intervention suggesting that people were calibrating their judgments correctly (Basol et al., 2021). Sometimes confidence is not impacted by the intervention at all (e.g., Harrop et al., 2023) and in one study, people became *less* confident and shared less content overall in the treatment group, which the authors interpreted as a sign of epistemic humility, which perhaps could be considered a desirable side effect (Lees et al., 2023).

A final and perhaps more complex discussion has erupted around the question of whether inoculation interventions mainly help people spot misleading content or whether they also boost correct “*discernment*” between true and false news. For example, an argument can be made that if inoculation interventions do not improve discernment, they could just be making people more skeptical of news media across the board (akin to an overactive immune system, attacking “healthy cells”). This is a complex question because the answer depends both on how we define discernment as well as misinformation. For example, consider that, although outright fake news is relatively uncommon in most people’s media diet, biased or misleading news is much more prevalent in both fake and sometimes credible outlets (Goel et al., 2023; van der Linden et al., 2023; Traber et al., 2023). Discernment is sometimes calculated as the difference between correct recognition of misleading techniques in the treatment versus the control group (Jigsaw, 2023) whereas others assume it to mean discernment between true and false item sets (e.g., Pennycook, 2023; Pennycook & Rand, 2019).

Moreover, whether the detection of fake news or improving discernment between “fake” and “real” news is important depends on the goal of the intervention (Guay et al., 2022). For example, consider that ‘discernment’ has historically not been a measurement feature of inoculation theory or the literature on misinformation more generally. Interventions that inoculate people against a particular conspiracy theory (e.g., 9/11) go on to

measure whether people are now more resistant to that specific conspiracy theory (Banas & Richards, 2017) not whether people can discern better between true and false conspiracy theories because that was not the aim of the inoculation intervention. In other words, the inoculation intervention does not help people spot real conspiracies or discern between the two, rather, its aim is to create resistance against false conspiracy theories. At population level, this seems reasonable as other interventions might focus on helping people identify credible news so not all interventions need to have the same goal (Guay et al., 2022).

However, one could certainly make the argument that if the intervention not only makes people more resistant to say, a conspiracy theory about 9/11, but also makes people less likely to believe in real news about 9/11, such heightened skepticism could be an undesirable side effect of the inoculation process. The reason why discernment is potentially interesting as an additional outcome measure is because it helps illustrate potential trade-offs of an intervention. For example, if people downgrade the accuracy of real news a little bit post-intervention (“false positives”) this could be acceptable as long as the boost in fake news recognition (“true positives”) is much higher so that discernment remains significant. But solely reporting on discernment can be misleading (Roozenbeek, Maertens et al., 2020), because discernment in and of itself does not illustrate how and why people are improving and whether the intervention is having the intended effect. For example, an intervention meant to increase people’s recognition of credible news can have significant discernment driven solely by helping people spot *false* news, which is still a good outcome, but theoretically inconsistent with the mechanisms and goals behind said intervention.

Modirrousta-Galian and Higham (2023) recently re-analyzed a selective number of active inoculation games using signal detection theory (SDT) with the specific aim of looking at how they impact both truth discernment (sensitivity) and response bias (a tendency to rate an item as true or false), which is important given that the literature on fake news has not sufficiently disentangled these two processes (Batailler et al., 2022). Although some of the included interventions, such as Roozenbeek and van der Linden (2019) and Basol et al., (2021, Study 1) showed significant discernment, there was no consistent effect across studies. Moreover, the authors argue that these interventions simply induce a conservative response bias where people become more skeptical of all news items (perhaps because the game draws attention to manipulation). First, while

interesting in its own right, it is important to note that these results differ from a wider systematic review and meta-analysis of our inoculation studies from Lu et al. (2023), who note that in contrast to Modirrousta-Galian and Higham (2023), our studies produced significant effects on truth discernment ($d = 0.20$, 95% CI; 0.13, 0.28, $p < 0.001$). Moreover, even in the studies analyzed by Modirrousta-Galian and Higham (2023), it seems premature to conclude that discernment suffers *because* of a response bias (skepticism), since sensitivity and response bias are technically independent of another, so there is some room for caution here as different psychological explanations can underpin each parameter (Batailler et al., 2022). It is also noteworthy that in Modirrousta-Galian and Higham's (2023) re-analysis, some experiments (e.g., Roozenbeek and van der Linden, 2019; Iyengar et al., 2022) and items sets (e.g., Set A-A in Roozenbeek et al., 2020) did not reveal evidence of response bias. Second, sometimes the heightened skepticism of real news is temporary (i.e., only evident on the immediate post-test), and in several cases the control groups also display response bias, which is peculiar as the control groups did not receive any information about manipulation. Another account for the response bias is therefore rooted in methodological artefacts where high ratios of fake to credible information in itself can induce response bias for studies with unbalanced items (Altay, Lyons, & Modirrousta-Galian, 2023). Modirrousta-Galian and Higham (2023) also acknowledge the importance of item effects and conclude that conservative response bias only seems to happen for "ambiguous" real news items. This makes some sense, as for items which are not clearly true—but instead more ambiguous in their presentation—people are more likely to err on the side of caution with a more conservative responding pattern (and perhaps rightly so in the absence of factual knowledge).

Second, it is relevant to note that in many of our earlier studies, only a few real news items were included because we did not intend to have a valid measure of discernment but were nonetheless interested in possible side effects. The psychometric reliability of the misinformation items is therefore generally high as they were specifically designed to measure the relevant misinformation techniques with enough items per technique to establish basic reliability. In contrast, the real news items were often exploratory and did not load onto a single factor nor did they have good reliability (we analyze and report them separately). In terms of discrimination, it is therefore worth pointing out that when you subtract a reliable item set from an unreliable one, the difference (discernment) is not

necessarily reliable so the discernment measure used by [Modirrousta-Galian and Higham \(2023\)](#) is of unknown quality.

In terms of response bias, this means that there is not enough opportunity for people to develop confidence about what a “true” signal looks like on these type of trials (especially if some of the true signals also contain hints of falsity) and we should generally be cautious about conclusions drawn from small stimuli sets ([Judd, Westfall, & Kenny, 2017](#)).

In fact, the third and much larger point to consider is whether Signal Detection Theory (SDT) is even an appropriate framework for evaluating inoculation interventions that are focused on degrees of manipulation rather than true-false binaries. SDT assumes a classification model where the true positive rate of a task can be plotted against the false positive rate (as part of an ROC curve). A classic example is the sensitivity of a medical test which classifies people as either having the disease or not ([Cook, 2008](#); [Cook, 2020](#)). But news classification does not work in the same manner: more often than not news is neither completely false nor entirely true and thus the categories are not mutually exclusive which violates the basic premise of Signal Detection Theory ([Szalma & Hancock, 2013](#)). Although recent calls for disentangling truth discernment from response bias can reveal important insights ([Batailler et al., 2022](#); [Modirrousta-Galian & Higham, 2023](#)), it is also key to consider whether the task at hand is one that concerns a true-false dichotomy versus the viewpoint that news presents itself on a continuum of bias and manipulation. For example, news can easily be true yet clearly misleading. Indeed, analyses shows that the use of emotional manipulation in “real” news (e.g., fear, outrage) has gone up by over 150% over the last few decades ([Rozado et al., 2022](#)). Thus, when a participant rates a “real” item as having lower reliability, say changing their rating from 6.9 to 6.7 (where 7 is very reliable) because the item contains a hint of polarization or emotional manipulation, this does not constitute a false positive (i.e., mistakenly viewing real news as “fake”) but instead is consistent with the purpose of the intervention: downgrading news that makes use of manipulation techniques regardless of source or intention. The point here is that SDT fails to capture the uncertainty inherent in real-world stimuli as an item can be classified as both correct detection and a false alarm depending on the degree to which the item represents the event of interest ([Murphy et al., 2004](#)). Interestingly, when real news items do not contain any manipulation and are obviously real, conservative responding patterns do not tend to occur in our interventions ([Modirrousta-Galian & Higham, 2023](#)). It may therefore well be the case

that “ambiguous” items are rated as somewhat less reliable by participants because they detect some form of manipulation. Because technique-based inoculation interventions do not teach people specific facts about the world, it seems entirely reasonable for people to become more skeptical when manipulation may be present in ambiguous headlines of unknown veracity. To mediate scholarly disagreement over what items represent the event of interest, one way to improve upon the analysis from [Modirrousta-Galian & Higham \(2023\)](#) is to conduct a fuzzy SDT analysis which would allow items to range on a continuum from 0 (for definitely neutral) to 1 (for clearly biased), maintaining greater nuance about the ambiguity of news items ([Szalma & Hancock, 2013](#)).

Overall, there is lively debate about the benefits of instilling a healthy dose of skepticism versus optimizing for truth discernment ([Batailler, Brannon, Teas, & Gawronski, 2022](#); [Roozenbeek & van der Linden, 2024](#)). I think it is important not to confuse trust in sources with skepticism toward specific headlines in a rating task. For example, it is probably desirable for people to retain a certain amount of skepticism toward questionable headlines from reliable outlets (“healthy skepticism”) as long as people do not lose trust in the source itself.¹⁰ In fact, it is not the case at all that people suddenly find real news “unreliable” in our research, merely slightly less but overall still highly reliable so the concern that people would suddenly stop believing in real news seems untenable, especially given the fact that conservative response bias is likely the result of a methodological artefact, i.e., being confronted with a high fake news to real news ratio in the testing phase (e.g., see [Altay, Lyons, & Modirrousta-Galian, 2023](#)).

A common concern is that because the base rate of fake news in the population is relatively low, people could end up making consequential false positive errors ([Modirrousta-Galian & Higham, 2023](#)).¹¹ Yet, we should also ask what the base rate is of highly objective and unbiased news in the population. If the base rate of unbiased news is also relatively low, then it is not necessarily a problem if people downgrade “real news” headlines for using various degrees of clickbait, sensationalism, polarization, emotion, or other manipulation techniques. It is also not clear how this type of skepticism might play out in social networks where belief

¹⁰ To date, there has been no empirical evidence of lower trust in mainstream sources and institutions following our interventions, though I encourage research to examine this possible side effect further.

¹¹ This seems unlikely in the interventions, because people still rate credible news as highly reliable both before and after playing the inoculation games in terms of the overall classing ([Maertens et al., 2020](#)).

consolidation takes place over time through exposure to conflicting information. To better understand this, we modeled the incidence rate of false positives in an agent-based social network where agents have prior beliefs and use cues to detect misinformation in three different kinds of information environments (mostly truthful news, mixed news, and mostly biased news) with the parameters tuned to effect-sizes from our inoculation interventions. At population level, we find that, although inoculation leads to slightly more false positive errors compared to controls, this happens regardless of the base rate of true or false information. The false positive rate also remains small compared to the number of correct rejections of false information (Pilditch et al., 2022). These results suggest that even in environments where most information is true, inoculation is unlikely to cause widespread errors.

Still, another concern is that even when people are successfully inoculated against manipulation techniques, they could selectively apply them to ideologically (in)congruent news. In other words, this could lead to partisan response bias by increasing the threshold for ideologically incongruent (fake) news (Gawronski, Ng, & Luke, 2023), which in turn could lead to polarization.¹² The evidence on ideology thus far suggests that the intervention is not moderated by ideology (Basol et al., 2020; Roozenbeek & van der Linden, 2019), reduces the perceived reliability of polarizing content, and evidence from people's choices in the game is mixed with respect to whether they craft ideologically congruent scenarios (Harrop et al., 2023). However, future research should disentangle the inoculation effect for ideologically concordant versus discordant headlines more specifically.¹³

Overall, non-significant discernment is only an issue if “suitable” item sets are used to assess the objective of the intervention. For example, there are various studies which show evidence of significant discernment in *Bad News*, including direct replications with local stimuli from India (Iyengar et al., 2022), item sets from other research groups (see Roozenbeek, van der Linden, & Nygren, 2021; Roozenbeek et al., 2022), as well as clear evidence of meta-analytic discernment (Lu et al., 2023). This indicates that

¹² In some ways, the original inoculation paradigm was already geared toward inoculating one attitude against conflicting views, which could result in increased polarity between the two positions.

¹³ It is also possible that inoculation can cause psychological reactance on some items, but current evidence seems to suggest that inoculation messages often help *mitigate* reactance (Clayton et al., 2023).

skepticism of real news might well be due to item or design effects rather than being a true causal impact of the intervention. In fact, we have previously shown that this effect is more likely to occur with uneven and non-balanced item sets (see [Experiment 3 in Maertens et al., 2020](#); [Roozenbeek, Maertens et al., 2020](#)). Indeed, later research that matched item sets to be balanced on either manipulative or non-manipulative characteristics found strong and consistent discernment effects ([Roozenbeek et al., 2022](#)). Meta-analyses also report how different research designs can influence the effect-size for discernment ([Lu et al., 2023](#)).

Relatedly, it is important to note that any negative impact on real news likely has little to do with the phenomenon of inoculation itself. For example, research finds that media literacy tips, fact-checking, and debunking all cause people to endorse false information less at the expense of making people somewhat more skeptical of “true” news ([Hameleers, 2023](#); [Hoes et al., 2023](#)). There is an important parallel here with the literature on the backfire effect: fact-checkers were led to believe for many years that fact-checking could lead people to reject true information even more ([Nyhan, 2021](#)), even though it was later shown that although backfire can occur for some extreme subpopulations, it is largely due to item reliability and design artefacts rather than a true feature of fact-checking and debunking (see [Swire-Thompson et al., 2020, 2022](#)). We therefore need to disentangle whether any observed skepticism of real news is due to design choices (or the items being selected) rather than being a true side effect of the intervention process itself. One way to do this is for future research to select balanced item sets of misleading false news, misleading true news, as well as neutral true news to identify whether unwarranted skepticism occurs for *neutral* true news.

Lastly, we have already started to explore how interventionists could potentially counter any undesirable side effects ([Leder et al., 2023](#)). One benefit of active inoculation games is that they are a dynamic “living” intervention, open to change so we can easily fine-tune different levels of preferred skepticism by optimizing for different outcomes. Maybe some heightened skepticism might actually be a beneficial side effect insofar it could lead people to investigate claims more thoroughly. Such hypotheses can be tested by altering the level of skepticism in the games. One novel way of doing this is by providing players with feedback. Feedback is a key mechanism behind adaptive learning in response to ambiguous stimuli because it reinforces “correct” responses and helps minimize prediction errors ([Butler et al., 2008](#); [Hattie & Timperley, 2007](#)), which should help

boost discernment outcomes. Indeed, across five recent studies (Leder et al., 2023), we implemented a simple feedback mechanism (already live in our interventions) to help players detect the manipulation techniques when present (versus not), which substantially increases discernment and AUC scores and mitigates response bias, mostly by changing how people evaluate real news (Table 1). Feedback could therefore be an important mechanism for researchers concerned with discernment scores and response bias.

6.4 Psychological herd immunity

It is somewhat remarkable that McGuire never took the concept of attitudinal inoculation to its logical conclusion: herd immunity (Compton, van der Linden et al., 2021; Lewandowsky & van der Linden, 2021; van der Linden et al., 2022; van der Linden, Maibach et al., 2017). Especially when considered from a social psychological perspective, the process of inoculation is fairly cognitive in its orientation. Yet, the way inoculation could be spread and passed onto others is entirely social. Indeed, research has found that after people receive an inoculation message they engage in so-called “post-inoculation talk” (PIT) where talking about the inoculation not only strengthens the initial inoculation (Ivanov et al., 2012) but also allows for the possibility of transfer via word-of-mouth in social networks (Compton & Pfau, 2009; van der Linden, Maibach et al., 2017). Interestingly, while the evidence for inoculation generating “talk” about a target issue is fairly well-established (e.g., see Ivanov et al., 2012, 2017), the exact process by which inoculation could spread in social networks has remained largely elusive and mostly theoretical (Compton & Pfau, 2009).

To try to further explore the potential for herd immunity in the context of misinformation, we created an agent-based model of a social network with belief-updating users (Pilditch et al., 2022). There are two types of agents in this model, citizens (who can choose to share information with others) or broadcasters (who broadcast various degrees of both truthful and misinformation). The disseminated information comes with varying misinformation ‘cues’ (such as conspiratorial reasoning or inflammatory language) that users can learn to detect given baseline sensitivity to these cues as well as via inoculation interventions. To make the simulations more realistic, the inoculation parameters were tuned according to the effect-sizes from our experiments in terms of how well people can recognize these techniques, their intentions to share misinformation with others, and the expected decay of the intervention over time. Across many simulations, we find that in the presence of biased information, users will self-select into

Table 1 Results from feedback experiment in *Bad News* (Leder et al., 2023).

Game	H condition	t	df	p	M _{diff}	Cohen's d	95% CI
Bad news							
Without feedback							
	Misinformation (Post)	-9.83	951	<.001	-0.375	-0.32	[-0.384, -0.253]
	Real news (Pre)	-5.95	951	<.001	-0.277	-0.19	[-0.257, -0.129]
	Discernment (Post)	1.99	951	0.047	0.098	0.06	[0.0009, 0.128]
With feedback							
	Misinformation (Post)	-6.96	943	<.001	-0.253	-0.23	[-0.291, -0.162]
	Real news (Pre)	4.75	943	<.001	0.206	0.16	[0.090, 0.219]
	Discernment (Post)	9.23	943	<.001	0.459	0.30	[0.235, 0.365]

echo chambers where beliefs become consolidated over time regardless of their veracity. In order to interrupt this process and maintain more “truthful” beliefs in the population—given different levels of inoculation, baseline sensitivity, and broadcasters—a critical threshold (60%) of the network needs to be inoculated in advance rather than in (20%) stages throughout to maximize effectiveness, suggesting it is important to front-load inoculation campaigns when considering the networked diffusion of misinformation (Pilditch et al., 2022).

In a yet unpublished PhD dissertation study (see Basol, 2022), we tried to experimentally manipulate post-inoculation talk to see if we could find empirical evidence for the notion that (1) individuals actually engage in PIT unprompted (which hadn’t been tested before), (2) whether they can vicariously inoculate others by passing on the critical content so that even individuals who were not “vaccinated” themselves can benefit from the inoculation, and (3) whether the spread of the “vaccine” can outpace the spread of “misinformation”. In two pre-registered, three-phase experiments we inoculated participants against misinformation about a fictitious chemical and recorded what information they would like to pass on to another subject in the form of a tweet. One week later, we asked whether participants had engaged in any talk about the inoculation themselves and to describe the content of the talk. We then used the messages people chose in the first phase as the inoculation message in a second experiment to model the efficacy of *vicarious* (“pass-along”) inoculation. We coded the open-ended talk responses according to a codebook which reflected key inoculation themes (e.g., forewarning-related talk, preemptive refutation material, counterarguments, etc. vs. repeating some form of the misinformation itself). We also distinguished between the quality of the talk itself, e.g., highly generic warnings versus more specific prebunks.

Although we found no changes in attitudes overall (likely because participants had never heard of the chemical in question), we did find that people in the inoculation condition reported greater post-inoculation talk, higher-quality talk, and greater self-reported resistance to the misinformation. Most importantly, inoculated individuals also passed on significantly less misinformation than those in the control conditions offering the first evidence that vicarious inoculation could reduce the spread of misinformation (Basol, 2023). Most research has looked at the role of PIT as a vocal component of inoculation in terms of its effect on the recipient, instead of examining the social diffusion of resistance. We were only able to look at a two-step sharing process here (i.e., what information participants chose to

pass on and using that as the inoculation for new participants in the second phase). Future research should try to model the diffusion of inoculation and misinformation in much longer social network chains to test the holy grail of inoculation theory; if enough people in a community are vaccinated, misinformation will no longer have a chance to spread.



7. Conclusion

The aim of the current chapter was to shed light on the possibility of inoculating people against misinformation by forewarning and exposing them to a weakened dose of the attacking material alongside persuasive refutations. Although inoculation theory has not seen much development in social psychology since the 1960's, significant theoretical and translational advancements have been made in the last decade(s), particularly in the context of combatting misinformation and propaganda in the real-world. Although the theory's focus has historically been restricted to "cultural truisms", I echo the important observation that the analogy was meant to be "more instructive, than prescriptive" (Compton et al., 2013, p. 233). Research has clearly evidenced that inoculation is not only possible in the context of controversial issues but that we should conceptually distinguish between *therapeutic* and *prophylactic* forms of inoculation depending on whether and how often people have already been exposed to potential misinformation. Further distinctions have been made between *passive* and *active* inoculation and *narrow-spectrum* and *broad-spectrum* immunity where participants can achieve resistance against a wide range of variants of the same misinformation technique, including material not mentioned in the initial inoculation. For the first time, inoculation has been scaled in real-world contexts including exposure to hundreds of millions of people on social media via interactive games, animated videos, and other novel campaigns, illustrating the scalability of this approach as it is being adopted by social media companies, public health authorities, and governments around the world. We have come to understand more about the key mechanisms behind inoculation effectiveness, including the role of motivation to defend oneself from misinformation (elicited by threat and forewarnings) and the role of ability in the form of refutational pre-emptions (or prebunks), and the need to boost these processes in people's memory to maintain immunity over time. Lastly, as with any intervention, side effects need to be documented, considered, and managed, including realistic expectations about

effect-sizes outside of laboratory conditions where interference is likely to be much greater. As attitudinal inoculation research is once again emerging as an important theory in social psychology and beyond, the most exciting days for inoculation are still ahead, including the possibility of achieving herd immunity to misinformation via social diffusion.

Acknowledgments

I would like to thank the UK government, the European Commission, the American Psychological Association (APA), the CDC, Google, Jigsaw, WhatsApp/Meta, the WHO, UN, the US State Department (GEC)/CISA, Jon Roozenbeek and the whole Cambridge Social Decision-Making Lab, DROG/TILT, Gusmanson Design and Lens Change/Studio You for their support and for making the research described in this chapter possible.

References

- Allen, J. N. L., Watts, D. J., & Rand, D. (2023). Quantifying the Impact of Misinformation and Vaccine-Skeptical Content on Facebook. <https://doi.org/10.31234/osf.io/nwsqa>.
- Altay, S., Nielsen, R. K., & Fletcher, R. (2022). Quantifying the “infodemic”: People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media*, 2, 1–29.
- Altay, S., Lyons, B., & Modirrousta-Galian, A. (2023, preprint). Exposure to higher rates of false news erodes media trust and fuels skepticism in news judgment. (<https://psyarxiv.com/t9r43>).
- Amazeen, M. A., Krishna, A., & Eschmann, R. (2022). Cutting the bunk: Comparing the solo and aggregate effects of prebunking and debunking COVID-19 vaccine misinformation. *Science Communication*, 44(4), 387–417.
- Badrinathan, S. (2021). Educative interventions to combat misinformation: Evidence from a field experiment in India. *American Political Science Review*, 115(4), 1325–1341.
- Badrinathan, S., & Chauchard, S. (2023). Researching and Countering Misinformation in the Global South. *Current Opinion in Psychology*, 101733.
- Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 6(10), 1372–1380.
- Ballew, M. T., Leiserowitz, A., Roser-Renouf, C., Rosenthal, S. A., Kotcher, J. E., Marlon, J. R., & Maibach, E. W. (2019). Climate change in the American mind: Data, tools, and trends. *Environment: Science and Policy for Sustainable Development*, 61(3), 4–18.
- Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Sadhana Pravin, M. (2019). WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. Department of Media and Communications, London School of Economics and Political Science. Available from (https://eprints.lse.ac.uk/104316/1/Banaji_whatsapp_vigilantes_exploration_of_citizen_reception_published.pdf).
- Banas, J., & Rains, S. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, 77(3), 281–311.
- Banas, J., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and meta-inoculation strategies. *Human Communication Research*, 39(2), 184–207.
- Banas, J., & 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.
- Basol, M. (2022). Essays on resistance against persuasion: Building, strengthening, and spreading attitudinal resistance through inoculation theory (Doctoral dissertation, University of Cambridge).

- 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), 1–9.
- Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data and Society*, 8(1), <https://doi.org/10.1177/205395172111013868>.
- Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, 17, 78–98.
- Benton, J. (2021). Facebook sent a ton of traffic to Chicago Tribune story. So why is everyone mad at them? *Nieman Lab*. (<https://www.niemanlab.org/2021/08/facebook-sent-a-ton-of-traffic-to-a-chicago-tribune-story-so-why-is-everyone-mad-at-them/>).
- Bertolotti, M., & Catellani, P. (2023). Counterfactual thinking as a prebunking strategy to contrast misinformation on COVID-19. *Journal of Experimental Social Psychology*, 104, 104404.
- Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2023). Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South. *Current Opinion in Psychology*, 101732.
- Blake, K. D., Willis, G., & Kaufman, A. (2020). Population prevalence and predictors of self-reported exposure to court-ordered, tobacco-related corrective statements. *Tobacco Control*, 29(5), 516–521.
- Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5), e2020043118.
- Bruns, H., Lewandowsky, S., Pennycook, G., Pantazi, M., Schmid, P., Krawczyk, M.W. ... Smillie, L. (2023, preprint). The role of (trust in) the source of prebunks and debunks of misinformation. Evidence from online experiments in four EU countries. <https://doi.org/10.31219/osf.io/vd5qt>.
- Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 918–928. <https://doi.org/10.1037/0278-7393.34.4.918>.
- Chan, M. P. S., & Albarracín, D. (2023). A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour*, 1–12.
- Chan, M. P. S., Jones, C. R., Jamieson, K. H., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11), 1531–1546.
- Chido-Amajuoyi, O. G., Yu, R. K., Agaku, I., & Shete, S. (2019). Exposure to court-ordered tobacco industry antismoking advertisements among US adults. *JAMA Network Open*, 2(7), e196935.
- Clayton, R. B., Compton, J., Reynolds-Tylus, T., Neumann, D., & Park, J. (2023). Revisiting the effects of an inoculation treatment on psychological reactance: A conceptual replication and extension with self-report and psychophysiological measures. *Human Communication Research*, 49(1), 104–111.
- Compton, J. (2005). Tracing the roots of resistance to influence: Comparison, contrast, and synthesis of Aristotelian rationality and inoculation. *STAM Journal*, 35, 1–23.
- Compton, J. (2013). Inoculation theory. In J.P. Dillard & L. Shen (Eds.), *The SAGE Handbook of Persuasion: Developments in theory and practice* (pp. 220–237).
- Compton, J. (2020). Prophylactic versus therapeutic inoculation treatments for resistance to influence. *Communication Theory*, 30(3), 330–343.
- Compton, J. (2021). Threat and/in inoculation theory. *International Journal of Communication*, 15(13), 4294–4306.

- Compton, J., & Pfau, M. (2009). Spreading inoculation: Inoculation, resistance to influence, and word-of-mouth communication. *Communication Theory, 19*, 9–28.
- Compton, J., & Ivanov, B. (2012). Untangling threat during inoculation–conferred resistance to influence. *Communication Reports, 25*(1), 1–13.
- Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass, 15*(6), e12602.
- Compton, J. A., & Pfau, M. (2005). Inoculation theory of resistance to influence at maturity: Recent progress in theory development and application and suggestions for future research. *Annals of the International Communication Association, 29*(1), 97–146.
- Cook, J. (2020). *Cranky Uncle vs climate change: How to understand and respond to climate science deniers*. New York, NY: Citadel Press.
- Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One, 12*(5), e0175799.
- Cook, J., Ellerton, P., & Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters, 13*(2) 024018.
- Cook, J., van der Linden, S., Maibach, E., & Lewandowsky, S. (2018). *The consensus handbook*. Available at: (<http://www.climatechangecommunication.org/all/consensus-handbook/>).
- Cook, J., Ecker, U. K. H., Trecek-King, M., Schade, G., Jeffers-Tracy, K., Fessmann, J., ... McDowell, J. (2022). The Cranky Uncle game—Combining humor and gamification to build student resilience against climate misinformation. *Environmental Education Research, 29*(4), 607–623.
- Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve. *Clinical Chemistry, 54*(1), 17–23.
- Cummings, K. M., Morley, C. P., & Hyland, A. (2002). Failed promises of the cigarette industry and its effect on consumer misperceptions about the health risks of smoking. *Tobacco Control, 11*(suppl 1), i110–i117.
- DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica, 90*(1), 81–116.
- Doosje, B., Moghaddam, F. M., Kruglanski, A. W., De Wolf, A., Mann, L., & Feddes, A. R. (2016). Terrorism, radicalization and de-radicalization. *Current Opinion in Psychology, 11*, 79–84.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich.
- Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology, 1*(1), 13–29. <https://doi.org/10.1038/s44159-021-00006-y>.
- Frieden, T. R., & Blakeman, D. E. (2005). The dirty dozen: 12 myths that undermine tobacco control. *American Journal of Public Health, 95*(9), 1500–1505.
- Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances, 7*(23), eabf1234.
- Gawronski, B., Ng, N. L., & Luke, D. M. (2023). Truth sensitivity and partisan bias in responses to misinformation. *Journal of Experimental Psychology: General, 152*, 2205–2236.
- Goel, P., Green, J., Lazer, D., & Resnik, P. (2023). Mainstream news articles co-shared with fake news buttress misinformation narratives. *arXiv preprint arXiv, 2308, 06459*.
- Goel, V., Raj, J., & Ravichandran, P. (July 18th, 2018). How WhatsApp leads mobs to murder in India. *The New York Times*. Available from (<https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html>).
- 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.

- Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior*, 21(2), 207–219.
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425), 374–378.
- Guay, B., Pennycook, G., & Rand, D. (2022, in press). How to think about whether misinformation interventions work. *Nature Human Behaviour*.
- Gunther, R., Beck, P. A., & Nisbet, E. C. (2019). “Fake news” and the defection of 2012 Obama voters in the 2016 presidential election. *Electoral Studies*, 61, 102030.
- Hameleers, M. (2023). The (un) intended consequences of emphasizing the threats of mis- and disinformation. *Media and Communication*, 11(2), 5–14.
- Harjani, T., Basol, M., Roozenbeek, J., & van der Linden, S. (2023). Gamified inoculation against misinformation in India: A randomised control trial. *Journal of Trial and Error*. <https://doi.org/10.36850/e12>.
- Harrop, I., Roozenbeek, J., Madsen, J., & van der Linden, S. (2023). Inoculation can reduce the perceived reliability of polarizing social media content. *International Journal of Communication*, 17, 5291–5315.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>.
- Hebel-Sela, S., Hameiri, B., & Halperin, E. (2022). The vicious cycle of violent intergroup conflicts and conspiracy theories. *Current Opinion in Psychology* 101422.
- Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2023). Prominent Misinformation Interventions Reduce Misperceptions but Increase Skepticism. <https://doi.org/10.31234/osf.io/zmpdu>.
- Ivanov, B., Rains, S. A., Geegan, S. A., Vos, S. C., Haarstad, N. D., & Parker, K. A. (2017). Beyond simple inoculation: Examining the persuasive value of inoculation for audiences with initially neutral or opposing attitudes. *Western Journal of Communication*, 81(1), 105–126.
- Ivanov, B., Rains, S. A., Dillingham, L. L., Parker, K. A., Geegan, S. A., & Barbati, J. L. (2022). The role of threat and counterarguing in therapeutic inoculation. *Southern Communication Journal*, 87(1), 15–27.
- Ivanov, B., Miller, C. H., Compton, J., Averbek, J. M., Harrison, K. J., Sims, J. D., & Parker, J. L. (2012). Effects of postinoculation talk on resistance to influence. *Journal of Communication*, 62(4), 701–718.
- Iyengar, A., Gupta, P., & Priya, N. (2022). Inoculation against conspiracy theories: A consumer side approach to India’s fake news problem. *Applied Cognitive Psychology*, 37(2), 290–303.
- Jiang, M., Gao, Q., & Zhuang, J. (2021). Reciprocal spreading and debunking processes of online misinformation: A new rumor spreading–debunking model with a case study. *Physica A: Statistical Mechanics and its Applications*, 565, 125572.
- Jigsaw (2023). *Defanging Disinformation’s Threat to Ukrainian Refugees*. Available from <https://medium.com/jigsaw/defanging-disinformations-threat-to-ukrainian-refugees-b164dbbc1c60>.
- Jigsaw (2023b). *Prebunking to build defenses against online manipulation tactics in Germany*. Available from <https://medium.com/jigsaw/prebunking-to-build-defenses-against-online-manipulation-tactics-in-germany-a1dbfbc67a1a>.
- Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, 47(8), 459–469.
- Jolley, D., & Paterson, J. L. (2020). Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology*, 59(3), 628–640.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, 68, 601–625.

- Koehler, D. J. (2016). Can journalistic “false balance” distort public perception of consensus in expert opinion? *Journal of Experimental Psychology: Applied*, 22(1), 24–38.
- Kostygina, G., Szczyпка, G., Tran, H., Binns, S., Emery, S. L., Vallone, D., & Hair, E. C. (2020). Exposure and reach of the US court-mandated corrective statements advertising campaign on broadcast and social media. *Tobacco Control*, 29(4), 420–424.
- Krause, N. M., Beets, B., Howell, E. L., Tosteson, H., & Scheufele, D. A. (2023). Collateral damage from debunking mRNA vaccine misinformation. *Vaccine*, 41(4), 922–929.
- Leder, J., Schellinger, L.V., Maertens, R., van der Linden, S., & Roozenbeek, J. (2023). Feedback boosts discernment and longevity for gamified misinformation interventions. [Manuscript under review].
- Lees, J., Banas, J. A., Linvill, D., Meirick, P. C., & Warren, P. (2023). The Spot the Troll Quiz game increases accuracy in discerning between real and inauthentic social media accounts. *PNAS nexus*, 2(4), pga094.
- Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2), 348–384.
- Lewandowsky, S., & Yesilada, M. (2021). Inoculating against the spread of Islamophobic and radical-Islamist disinformation. *Cognitive Research: Principles and Implications*, 6, 1–15.
- 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.
- Lewandowsky, S., Stritzke, W. G., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. *Psychological Science*, 16(3), 190–195.
- Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131.
- Lewandowsky, S., Cook, J., Ecker, U., Albarracín, D., Kendeou, P., Newman, E. J., & Zaragoza, M. S. (2020). The debunking handbook 2020. (<https://www.climatechangecommunication.org/wp-content/uploads/2020/10/DebunkingHandbook2020.pdf>).
- Loomba, S., De Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3), 337–348.
- Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X. D. (2023, in press). Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *Journal of Medical Internet Research*. <https://doi.org/10.2196/49255>. <https://pubmed.ncbi.nlm.nih.gov/37560816/>.
- Lumsdaine, A. A., & Janis, I. L. (1953). Resistance to “counterpropaganda” produced by one-sided and two-sided “propaganda” presentations. *Public Opinion Quarterly*, 17(3), 311–318.
- Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118(23), e2019527118.
- Ma, J., Chen, Y., Zhu, H., & Gan, Y. (2023). Fighting COVID-19 misinformation through an online game based on the inoculation theory: Analyzing the mediating effects of perceived threat and persuasion knowledge. *International Journal of Environmental Research and Public Health*, 20(2), 980.
- Maertens, R. (2023). The long-term effectiveness of inoculation against misinformation: An integrated theory of memory, threat, and motivation (Doctoral dissertation, University of Cambridge).
- Maertens, R., Anseel, F., & van der Linden, S. (2020). Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology*, 70, 101455.

- Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2020). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27(1), 1–16.
- Maertens, R., Roozenbeek, J., Simons, J., Lewandowsky, S., Maturo, V., Goldberg, B., ... van der Linden, S. (2023). Psychological booster shots targeting memory increase long-term resistance against misinformation. <https://doi.org/10.31234/osf.io/6r9as>.
- Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., & van der Linden, S. (2023). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, 1–37.
- McGuire, W. J. (1961a). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Sociometry*, 24(2), 184–197. <https://doi.org/10.2307/2786067>.
- McGuire, W. J. (1961b). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *The Journal of Abnormal and Social Psychology*, 63(2), 326–332. <https://doi.org/10.1037/h0048344>.
- McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology*, 1, 191–229.
- McGuire, W. J. (1970). A vaccine for brainwash. *Psychology Today*, 3(9), 37–64.
- 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.
- McGuire, W. J., & Papageorgis, D. (1962). Effectiveness of forewarning in developing resistance to persuasion. *Public Opinion Quarterly*, 26(1), 24–34. <https://doi.org/10.1086/267068>.
- McPhedran, R., Ratajczak, M., Mawby, M., King, E., Yang, Y., & Gold, N. (2023). Psychological inoculation protects against the social media infodemic. *Scientific Reports*, 13(1), 5780.
- 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, 2211–2237.
- Morgan, J. C., & Cappella, J. N. (2023). The effect of repetition on the perceived truth of tobacco-related health misinformation among US adults. *Journal of Health Communication*, 28(3), 182–189.
- Motta, M., & Stecula, D. (2021). Quantifying the effect of Wakefield et al. (1998) on skepticism about MMR vaccine safety in the US. *PLoS One*, 16(8), e0256395.
- Murphy, L., Szalma, J. L., & Hancock, P. A. (2004). *Comparison of fuzzy signal detection and traditional signal detection theory: Analysis of duration discrimination of brief light flashes. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 48*, Sage CA: Los Angeles, CA: SAGE Publications, 2494–2498.
- Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus’ forgetting curve. *PLoS one*, 10(7), e0120644.
- 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).
- Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15).
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303–330.
- O’Mahony, C., Brassil, M., Murphy, G., & Linehan, C. (2023). The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLoS One*, 18(4), e0280902.

- Oreskes, N., & Conway, E. (2010). *Merchants of doubt*. London: Bloomsbury.
- Osborn, W. W. (1939). An experiment in teaching resistance to propaganda. *The Journal of Experimental Education*, 8(1), 1–17.
- Parker, K. A., Rains, S. A., & Ivanov, B. (2016). Examining the “blanket of protection” conferred by inoculation: The effects of inoculation messages on the cross-protection of related attitudes. *Communication Monographs*, 83(1), 49–68.
- Pennycook, G. (2023). A framework for understanding reasoning errors: From fake news to climate change and beyond. *Advances in Experimental Social Psychology*, 67, 131–208.
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50.
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7), 770–780.
- Petty, R. E., & Cacioppo, J. T. (1977). Forewarning, cognitive responding, and resistance to persuasion. *Journal of Personality and Social Psychology*, 35(9), 645–655.
- Pfau, M. (1997). Progress in communication sciences: Advances in persuasion. In G. A. Barnett, & F. J. Boster (Vol. Eds.), *Inoculation model of resistance to influence*. 13. *Inoculation model of resistance to influence* (pp. 133–171). Greenwich, CT: Ablex.
- Pfau, M., Ivanov, B., Houston, B., Haigh, M., Sims, J., Gilchrist, E., ... Richert, N. (2005). Inoculation and mental processing: The instrumental role of associative networks in the process of resistance to counterattitudinal influence. *Communication Monographs*, 72(4), 414–441. <https://doi.org/10.1080/03637750500322578>.
- Pierrri, F., Perry, B. L., DeVerna, M. R., Yang, K. C., Flammini, A., Menczer, F., & Bryden, J. (2022). Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports*, 12(1), 5966.
- Pilditch, T. D., Roozenbeek, J., Madsen, J. K., & van der Linden, S. (2022). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science*, 9(8), 211953.
- Piltch-Loeb, R., Su, M., Testa, M., Goldberg, B., Braddock, K., Miller-Idriss, C., Maturo, V., & Savoia, E. (2022). Testing the efficacy of attitudinal inoculation videos to enhance COVID-19 vaccine acceptance: A quasi-experimental intervention trial. *JMIR Public Health and Surveillance*, 8(6), <https://doi.org/10.2196/34615>.
- Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. e2104235118 *Proceedings of the National Academy of Sciences*, 118(37), <https://doi.org/10.1073/pnas.2104235118>.
- Pryor, B., & Steinfatt, T. M. (1978). The effects of initial belief level on inoculation theory and its proposed mechanisms. *Human Communication Research*, 4(3), 217–230.
- Readfearn, G. (2016). Revealed: Most popular climate story on social media told half a million people that science was a hoax. *Desmog*. Available from (<https://www.desmogblog.com/2016/11/29/revealed-most-popular-climate-story-social-media-told-half-million-people-science-was-hoax>).
- Rędzio, A. M., Izydorzak, K., Muniak, P., Kulesza, W., & Doliński, D. (2023). Is the COVID-19 bad news game good news? Testing whether creating and disseminating fake news about vaccines in a computer game reduces people’s belief in anti-vaccine arguments. *Acta Psychologica*, 236, 103930.
- Roozenbeek, J., & van der Linden, S. (2018). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5), 570–580.
- Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Nature Humanities and Social Sciences Communications*, 5, 65. <https://doi.org/10.1057/s41599-019-0279-9>.

- Roozenbeek, J., & van der Linden, S. (2020). Breaking Harmony Square: A game that “inoculates” against political misinformation. *The Harvard Kennedy School (HKS) Misinformation Review*, 1(8), <https://doi.org/10.37016/mr-2020-47>.
- Roozenbeek, J., & van der Linden, S. (2024). *The psychology of misinformation*. Cambridge, UK: Cambridge University Press.
- Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on “inoculation” theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 1, 2.
- Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy nudges? A pre-registered direct replication of Pennycook et al. (2020). *Psychological Science*, 32(7), 1–10. (<https://doi.org/10.1177/09567976211024535>).
- Roozenbeek, J., Traber, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, 9(5), 211719. <https://doi.org/10.1098/rsos.211719>.
- Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering misinformation: Evidence, knowledge gaps, and implications of current interventions. *European Psychologist*. <https://doi.org/10.31234/osf.io/b52um>.
- Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2020). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educational and Psychological Measurement*, 81(2), 340–362.
- 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.
- Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(10), 201199.
- Rozado, D., Hughes, R., & Halberstadt, J. (2022). Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLoS One*, 17(10), e0276367.
- Sagarin, B. J., Cialdini, R. B., Rice, W. E., & Serna, S. B. (2002). Dispelling the illusion of invulnerability: The motivations and mechanisms of resistance to persuasion. *Journal of Personality and Social Psychology*, 83(3), 526–541. <https://doi.org/10.1037/0022-3514.83.3.526>.
- Saleh, N.F., Makki, F., van der Linden, S., & Roozenbeek, J. (2023, accepted). Inoculating against extremist persuasion techniques – Results from a randomised controlled trial in Post-Conflict Areas in Iraq. *Advances in Psychology*.
- Saleh, N. F., Roozenbeek, J., Makki, F. A., McClanahan, W. P., & van der Linden, S. (2021). Active inoculation boosts attitudinal resistance against extremist persuasion techniques: A novel approach towards the prevention of violent extremism. *Behavioural Public Policy*, 1–24.
- Schmid-Petri, H., & Bürger, M. (2022). The effect of misinformation and inoculation: Replication of an experiment on the effect of false experts in the context of climate change communication. *Public Understanding of Science*, 31(2), 152–167.
- Schubatzky, T., & Haagen-Schützenhöfer, C. (2023). Inoculating adolescents against climate change misinformation. In *Fostering Scientific Citizenship in an Uncertain World: Selected Papers from the ESERA 2021 Conference* (pp. 275–292). Cham: Springer International Publishing.
- Seifert, C. M. (2002). *The continued influence of misinformation in memory: What makes a correction effective?* *Psychology of learning and motivation*, Vol. 41, Academic Press, 265–292.
- Swire-Thompson, B., & Lazer, D. (2019). Public health and online misinformation: Challenges and recommendations. *Annual Review of Public Health*, 41, 433–451.

- Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, 9(3), 286–299.
- Swire-Thompson, B., Miklaucic, N., Wihbey, J., Lazer, D., & DeGutis, J. (2022). Backfire effects after correcting misinformation are strongly associated with reliability. *Journal of Experimental Psychology: General*, 151(7), 1655–1665. <https://doi.org/10.1037/xge0001131>.
- Szalma, J. L., & Hancock, P. A. (2013). A signal improvement to Signal Detection Analysis: Fuzzy SDT on the ROCs. *Journal of Experimental Psychology: Human Perception and Performance*, 39(6), 1741–1762.
- Tormala, Z. L., & Petty, R. E. (2004). Source credibility and attitude certainty: A meta-cognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, 14(4), 427–442.
- Traberg, C. S. (2022). Misinformation: Broaden definition to curb its societal influence. *Nature*, 606(7915), 653.
- Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, 185, 111269.
- Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The Annals of the American Academy of Political and Social Science*, 700(1), 136–151.
- Traberg, C. S., Harjani, T., Basol, M., Biddlestone, M., Maertens, R., Roozenbeek, J., & van der Linden, S. (2023). *Prebunking against misinformation in the modern digital age. Managing infodemics in the 21st century*. Cham: Springer, 99–111.
- van der Linden, S., Albarraçin, D., Fazio, L. K., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. J. (2023). *Using psychological science to understand and fight health misinformation: An APA consensus statement*. Washington, DC: American Psychological Association. <https://www.apa.org/pubs/reports/health-misinformation>.
- van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3), 460–467.
- van der Linden, S. (2023a). *Foolproof: Why misinformation infects our minds and how to build immunity*. New York, NY: WW Norton.
- van der Linden, S. (2023b). We need a gold standard for randomised control trials studying misinformation and vaccine hesitancy on social media. *BMJ*, 381, 1007.
- van der Linden, S., & Roozenbeek, J. (2022). A psychological “vaccine” against fake news: From the lab to worldwide implementation. In N. Mažar, & D. Soman (Eds). *Behavioral Science in the Wild*. University of Toronto Press.
- van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2), 1600008.
- van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., & Lewandowsky, S. (2017). Inoculating against misinformation. *Science*, 358(6367), 1141–1142.
- Vivion, M., Anassour Laouan Sidi, E., Betsch, C., Dionne, M., Dubé, E., & Driedger, S. M. Canadian Immunization Research Network (CIRN). (2022). Prebunking messaging to inoculate against COVID-19 vaccine misinformation: An effective strategy for public health. *Journal of Communication in Healthcare*, 15(3), 232–242.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
- Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1), 136–144.
- Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3), 423–441.

- Walter, N., & Tukachinsky, R. (2019). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, 47(2), 155–177.
- Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3), 350–375.
- Wardle, C. (2018). The need for smarter definitions and practical, timely empirical research on information disorder. *Digital Journalism*, 6(8), 951–963.
- WEF. (2022). *The Global Risks Report 2022*. World Economic Forum. Available from (https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2022.pdf).
- West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15), e1912444117.
- Williams, M. N., & Bond, C. M. (2020). A preregistered replication of “Inoculating the public against misinformation about climate change”. *Journal of Environmental Psychology*, 70, 101456.
- Wilson, S. L., & Wiysonge, C. (2020). Social media and vaccine hesitancy. *BMJ Global Health*, 5(10), e004206.
- Wood, M. L. (2007). Rethinking the inoculation analogy: Effects on subjects with differing preexisting attitudes. *Human Communication Research*, 33(3), 357–378.
- Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes’ steadfast factual adherence. *Political Behavior*, 41, 135–163.
- Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, 395(10225), 676. [https://doi.org/10.1016/S0140-6736\(20\)30461-X](https://doi.org/10.1016/S0140-6736(20)30461-X).