Countering Misinformation and Fake News Through Inoculation and Prebunking

Stephan Lewandowsky[1] & Sander van der Linden[2]

[1] University of Bristol and University of Western Australia

[2] Department of Psychology, University of Cambridge

Author note

Both authors contributed equally.

Correspondence concerning this article should be addressed to Stephan Lewandowsky, School of Psychological Science, 12A Priory Road, University of Bristol, Bristol BS8 1TU. E-mail: stephan.lewandowsky@bristol.ac.uk

Abstract

There has been increasing concern with the growing infusion of misinformation, or "fake news", into public discourse and politics in many western democracies. Our article first briefly reviews the current state of the literature on conventional countermeasures to misinformation. We then explore proactive measures to prevent misinformation from finding traction in the first place that are based on the psychological theory of "inoculation". Inoculation rests on the idea that if people are forewarned that they might be misinformed and are exposed to weakened examples of the ways in which they might be misled, they will become more immune to misinformation. We review a number of techniques that can boost people's resilience to misinformation, ranging from general warnings to more specific instructions about misleading (rhetorical) techniques. We show that based on the available evidence, inoculation appears to be a promising avenue to help protect people from misinformation and "fake news".

Keywords; Fake News, Misinformation, Inoculation Theory, Prebunking.

Countering Misinformation and Fake News Through Inoculation and Prebunking

"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)

"Just remember, what you're seeing and what you're reading is not what's happening." (U.S. President Donald Trump, 24 July 2018)

"Post-truth" was nominated word of the year by *Oxford* dictionaries in 2016, to describe "circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief" (OED, 2016). Two political events in 2016 triggered the concern with truth–or rather its absence: The Brexit referendum in the U.K. and the election of Donald Trump in the U.S. During the Brexit referendum, the public's "epistemic rights"—that is, their right to be adequately informed—were serially violated by the British tabloids (Watson, 2018), and during the U.S. presidential campaign, independent fact checkers judged 70% of all statements by Donald Trump to be false or mostly false.

This situation invites at least two questions: First, can "fact-checking" provide a solution to "post-truth" politics? Second, instead of solely relying on fact-checking, could the public be given the skills and tools required to manage an environment in which political misinformation abounds?

Misinformation and Society

Misinformation sticks. Erasing "fake news" from one's memory is a challenging task, even under the best of circumstances; that is, in the psychological laboratory when participants

are motivated to be accurate and are free from distraction (for a review, see Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). In the cardinal misinformation experiment, people are presented with a fictitious scripted story (e.g., about a warehouse fire). In one condition, information that was presented early on (e.g., that oil paint had been found in a wiring cabinet) is explicitly corrected later in the script (e.g., the wiring cabinet was actually empty). In a control condition, the script never contains a correction and the wiring cabinet is presented as empty from the outset (e.g., Ecker, Lewandowsky, Swire, & Chang, 2011; Johnson & Seifert, 1994; Wilkes & Leatherbarrow, 1988). Although most participants can recall the correction, when present, after they have finished processing the script, they continue to rely on the original misinformation on an inference test. That is, when asked to explain why there was "so much black smoke", participants might refer to oil paint in the wiring cabinet. This "continued influence effect" of misinformation has been demonstrated repeatedly (for reviews, see Chan, Jones, Jamieson, & Albarracín, 2017; Lewandowsky et al., 2012; Swire & Ecker, 2017).

Continued influence of political misinformation

When circumstances are less controlled than in the laboratory, as in most real-life political events involving complex and messy events, false memories for non-existent events can be strikingly frequent. For example, Murphy, Loftus, Grady, Levine, and Greene (2019) presented participants in Ireland with true and false news stories relating to the referendum on abortion in the Republic of Ireland. Participants correctly recognized the true stories 56% of the time, but they also reported a distinct memory for one of the fabricated stories (invented by the experimenters) 37% of the time. Qualitative responses suggested that some participants reported rich and detailed false memories for one of the fabricated events. It is therefore perhaps unsurprising that the persistence of political misinformation can take on epic proportions. To

illustrate, consider the mythical Weapons of Mass Destruction (WMDs) that were alleged to be in Iraq and that were cited as the reason for the invasion of 2003. The constant drumbeat of "WMD, WMD, WMD" in the media and among politicians in the lead-up to the invasion, followed by innumerable media reports of "preliminary tests" that tested positive for chemical weapons during the early stages of the conflict—but ultimately were never confirmed by more thorough follow-up tests—created a strong impression that those weapons had been discovered. This impression was so powerful that notable segments of the American public continued to believe, up until at least 2014, that either the U.S. had found WMDs in Iraq or that Iraq had hidden the weapons so well that they escaped detection. Jacobson (2010) reviewed polling data from 2006 through 2009 and found that around 60% of Republicans (and around 20% of Democrats) believed in the existence of Iraqi WMDs, with little evidence of a decline of those false beliefs over time. A poll from December 2014 pegged erroneous beliefs in WMDs at 51% for Republicans and 32% for Democrats (http://publicmind.fdu.edu/2015/false/), confirming the longevity of those false beliefs. Mistaken beliefs in WMD thus persisted for around a decade after the *absence* of WMDs in Iraq had become the official U.S. position with the Duelfer report (September 2004; https://www.cia.gov/library/reports/general-reports-1/iraq_wmd_2004/).

Persistent false beliefs in war-related information were also observed with specific events during the initial stages of the invasion of Iraq. In a study conducted before the Marines reached Baghdad, Lewandowsky, Stritzke, Oberauer, and Morales (2005) presented participants with specific war-related items from the news media, some of which had been subsequently corrected. Participants were asked to indicate their belief in the items, as well as their recollection of the original information and memory for its correction. Among U.S. participants, even those individuals who were certain that the information had been retracted, continued to believe it to be true (Lewandowsky et al., 2005). The ironic co-existence of acknowledgement of a correction ("I

know that X is false") and continued belief ("I believe X to be true") or reliance ("I act like I believe X") on discredited information are hallmarks of the cognitive fallout from misinformation in the political arena. This fallout can manifest itself in a number of different ways.

Corrections of falsehoods but not feelings

There are repeated demonstrations that people can update their specific factual beliefs in response to corrections, but that those changes in belief have no politically-relevant downstream consequences, such as affecting voting intentions and favorability ratings of a candidate. In an experiment conducted during the U.S. primary campaign in 2016, Swire et al. (2017) presented more than 2,000 online participants with statements made by Donald Trump on the campaign trail. Half the statements shown to participants were true (e.g., "the U.S. spent $2 trillion on the war in Iraq") and the other half consisted of false claims (e.g., "vaccines cause autism"). Participants rated their belief in those statements (from "definitely false" to "definitely true"). Participants were then presented with corrections of the false statements and affirmations of the correct statements. On a subsequent test, belief ratings changed according to the experimental intervention: All participants, including Trump supporters, believed statements less after they were identified as false, and they believed them more after they were affirmed as being correct. However, for Trump supporters there was no association between the extent to which they shifted their belief when a statement was corrected and their feelings for Trump or their intention to vote for him. Thus, it seems that Trump's false claims did not matter to his supporters—at least they did not matter sufficiently to alter their feelings or voting intentions.

The same result was obtained in a study by Nyhan, Porter, Reifler, and Wood (2019) using a different methodology. They presented participants with a single incorrect claim made by Donald Trump (about crime rates), which was followed by various different types of correction

and a single belief rating. Trump supporters again showed that they were sensitive to the corrections, in comparison to a no-correction control condition. However, just as in the study by Swire et al. (2017), the correction had no effect on participants' favorability ratings of Donald Trump.

The basic pattern of results was replicated by Swire-Thompson, Ecker, Lewandowsky, and Berinsky (2020) in a study that also included supporters of Bernie Sanders and statements by Sanders (in addition to Trump supporters and statements by Trump). Supporters of both politicians adjusted beliefs in statements after being told they were false (or true), but those corrections typically did not affect their support for their favored candidate. It was only when there were four times as many false statements as true statements attributed to Trump or Sanders, that a statistically significant decline in support for the candidate was observed, although the effect size was small. (There were also small differences between supporters of Sanders and Trump but they are not relevant here.)

The persistent support for a politician even after he has been shown to make numerous false claims meshes well with public-opinion data about partisans' perceptions of President Trump. An NBC poll conducted in April 2018 revealed that 76% of Republicans thought that President Trump tells the truth "all or most of the time" (Arenge, Lapinski, & Tallevi, 2018). By contrast, only 5% of Democrats held that view. Essentially the same pattern was obtained by a Quinnipiac University poll in November 2018 (Quinnipiac, 2018).

The fallout from misinformation

It takes little imagination to realize that misinformed individuals are unlikely to make optimal decisions, and that even putting aside one's political preferences, this can have adverse

consequences for society as a whole. For example, following the unsubstantiated—and now thoroughly debunked (DeStefano & Thompson, 2004; Godlee, Smith, & Marcovitch, 2011)—claim of a link between childhood vaccinations and autism, numerous parents (largely in the U.K.) decided not to immunize their children. These misinformation-driven choices led to a marked increase in vaccine-preventable diseases, and substantial effort and expenditure were required to resolve this public-health crisis (Larson, Cooper, Eskola, Katz, & Ratzan, 2011; Poland & Spier, 2010).

Misinformation has also become associated with acts of violence or vandalism. In Myanmar, the military orchestrated a propaganda campaign on Facebook that targeted the country's Muslim Rohingya minority group. The ensuing violence forced 700,000 people to flee the country (Mozur, 2018). Violence can also arise without a directed campaign: In India, false rumours about child kidnappers shared via WhatsApp incited at least 16 mob lynchings in 2018, leading to the deaths of 29 innocent people (Dixit & Mac, 2018).

And at the time of this writing, the worldwide COVID-19 pandemic has given rise to multiple conspiracy theories and misleading news stories that have found considerable traction, with adverse consequences for society (van der Linden, Roozenbeek, & Compton, 2020). For example, 29% of Americans believe that COVID-19 was created in a laboratory (Schaeffer, 2020). In the U.K. the belief that 5G mobile technology is associated with COVID-19 has led to vandalism of infrastructure, with numerous cellphone masts being set alight by arsonists (Lewandowsky & Cook, 2020). About one quarter of the British public consistently endorses some form of conspiracy related to COVID-19 (Freeman, Waite, et al., 2020; see also Brennen, Simon, Howard, & Nielsen, 2020; Roozenbeek et al., 2020a). There is currently widespread concern among public-health officials that disinformation campaigns may curtail uptake of a

COVID-19 vaccine, which at the time of this writing is being rolled out in the U.K. (Peretti-Watel et al., 2020). Although acceptance of the new vaccine is high in the U.K. as of November 2020 (Freeman, Loe, et al., 2020), the trend in acceptance in many countries at the end of 2020 has been downward (Babalola et al., 2020). Belief in COVID-19 related misinformation has in fact been linked to reduced compliance with public health guidelines and lower reported willingness to take the vaccine and recommend it to others (Roozenbeek et al., 2020a).

The toxic fallout from misinformation is not limited to those direct consequences. Other more insidiuous fallouts may involve people's reluctance to believe in facts altogether. There have been numerous demonstrations that the presence of misinformation undermines the effects of accurate information. In one study, van der Linden, Leiserowitz, Rosenthal, and Maibach (2017) showed that when participants were presented with both a persuasive fact and a related piece of misinformation, belief overall was unaffected—the misinformation cancelled out the fact. McCright, Charters, Dentzman, and Dietz (2016) found that the presence of a contrarian counter frame cancelled out valid climate information, and the same effect was also observed by Cook, Lewandowsky, and Ecker (2017).

Misinformation does not just misinform. It also undermines democracy by calling into question the knowability of information altogether. And without knowable information deliberative democratic discourse becomes impossible (for an elaboration of those concerns, see Lewandowsky et al., 2017b; Lewandowsky et al., 2017a). Fortunately, we are not entirely powerless in confronting the "post-truth" malaise.

Confronting the "post-truth" world

Debunking of misinformation

Although the effectiveness of corrections in general is often debated, there is broad agreement in the literature that corrections of misinformation are more likely to be successful if the correction is accompanied by an alternative explanation, or if suspicion is aroused over the initial source of the misinformation (e.g., Lewandowsky et al., 2012). That is, telling people that negligence was *not* a factor in a story about a fictitious warehouse fire (i.e., stating that a wiring cabinet was empty after negligence was first implied by claiming it contained oil paint) is insufficient for participants to dismiss that information. Telling people instead that arson, rather than negligence, was to blame for the fire (by referring to petrol-soaked rags that were found at the scene), successfully eliminates reliance on the initial misinformation (e.g., Ecker, Lewandowsky, & Tang, 2010; Ecker, Lewandowsky, Cheung, & Maybery, 2015; Johnson & Seifert, 1994).

If a clear causal alternative is not available—as, for example, when attempting to rebut conspiracy theories about the disappearance of Malaysian Airlines flight MH370 over the Indian Ocean (MacLeod, Winter, & Gray, 2014)—arousing suspicion about the source of misinformation may be another technique to achieve debunking. For example, when mock jurors are admonished to disregard tainted evidence presented when reaching a verdict during a mock trial, they demonstrably continue to rely on that tainted evidence, similar to the way in which the oil paint in the wiring cabinet continues to affect participants' reasoning even when the cabinet was actually empty. Reliance on tainted evidence disappears only when jurors are made suspicious of the motives underlying the dissemination of the tainted evidence in the first place,

for example because it may have been planted by the prosecutor's office (Fein, McCloskey, & Tomlinson, 1997).

Although the success of these routes to debunking has been repeatedly established in the laboratory, their applicability outside the laboratory "in the wild" encounters at least three distinct problems. First, a causal alternative can only be effective to the extent that it exists or that it is accepted. There are many situations in which an alternative explanation may be unknown even though it is clear that the original information is false. For example, the claim that Malaysian Airlines flight MH370 was abducted by space aliens can be confidently identified as false; however, no well-established causal alternative exists that could be used to replace that claim.

Turning to the second problem, in other circumstances a causal alternative may exist, but it may come with ideological or political baggage that prevents some people from accepting it. The same problem also arises when skepticism of the source of misinformation is advisable: even though there may be good reasons to question the motives or credibility of a source, these reasons may not necessarily be accepted by the target audience. This problem can be illustrated with a study by Lewandowsky et al. (2005), which probed the public's knowledge and belief in war-related events during the early stages of the invasion of Iraq in 2003. Participants were presented with news items that had either been corrected by official sources after they were published or were thought to be true at the time. Participants were first asked for their belief in the items and whether they had heard of them previously, before being presented with a correction (where it existed) and a second set of belief ratings. Lewandowsky et al. (2005) found that people who accepted as true the official *casus belli*, namely the elimination of Weapons of Mass Destruction (WMD) thought to be hidden in Iraq, were likely to believe in news reports that *they knew had been corrected*. Those participants thus exhibited the quintessential ironic attribute of the

continued influence effect: knowledge that a piece of information is false accompanied by continued belief. By contrast, people who were skeptical of the official reason for the war, and who thought it was initiated over something other than WMD, were better able to dismiss false information and accept true statements. On the one hand these results affirm the importance of skepticism and its benefits to processing of information about contested events. On the other hand, given the highly partisan landscape of public opinion surrounding Iraq, with many Republicans—and considerably fewer Democrats—continuing to (mistakenly) believe that Iraq possessed WMDs in 2003 (e.g., Jacobson, 2010; Kull, Stephens, Weber, Lewis, & Hadfield, 2006), and with that belief being strongly associated with endorsement of the war (Kull et al., 2006), it is unlikely that provision of an alternative cause of the war would have been accepted by partisan supporters of the Bush administration's decision to invade. Indeed, Nyhan and Reifler (2010) showed that under certain circumstances a corrective message about WMDs can lead to an ironic further entrenchment of Republicans' false beliefs.[1]

---

[1] The "backfire" effect reported by Nyhan and Reifler (2010) has been found to be less common than initially thought (Guess & Coppock, 2018; Wood & Porter, 2018). We are therefore reluctant to expect backfire effects generally; however, the exact replication of Nyhan and Reifler (2010) reported by Wood and Porter (2018) (their Figure 5) are visually identical to those reported by Nyhan and Reifler (2010). When corrections challenge worldviews, we should therefore still be sensitive to the possibility of a backfire effect even though we should not routinely expect it.

The same problem continues to affect contemporary American public discourse. In light of repeated surveys showing that Republicans consider President Trump to be honest, the extensive archive of his misleading and false statements that is being accumulated by the *Washington Post's* fact-checker database is unlikely to convince supporters that Donald Trump's trustworthiness may be questionable. Conversely, Trump supporters may well question the accuracy of mainstream media such as the *Washington Post* that Trump is consistently dismissing as sources of "fake news" or even "enemies of the people". Under those circumstances, skepticism is likely to be driven more by partisan motivations than concern about the relevant evidence. In support, a recent study by van der Linden, Panagopoulos, and Roozenbeek (2020) found that the first media association that comes to mind for Republicans when they hear the phrase "fake news" is "CNN". CNN has been a frequent target of the President's ire. Under these circumstances, it is not entirely surprising that corrections fail to alter people's feelings about their preferred candidate (Swire et al., 2017; Swire-Thompson et al., 2020).

A final problem with debunking is that it is often forced to adapt a disadvantageous framing at a disadvantageous time. One often unavoidable attribute of corrections is that they tacitly accept someone else's rhetorial framing, thereby permitting the actor who promulgated the original falsehood to set the agenda. For example, a government official who announces that there are "no plans for a carbon tax" in response to a newspaper article falsely hinting at a tax may achieve a reduction in the specific belief that a carbon tax is imminent. However, the correction is keeping the concept of a "carbon tax" in the public realm, possibly deflecting public attention away from the government's actual agenda. The continued mention of a "carbon tax" may have additional fallout, for example by making people who oppose new taxes think about climate change mitigation as a greater threat than climate change itself—notwithstanding the fact that the climate crisis is now considered an acute emergency by many scientists (e.g., Gills &

Morgan, 2020) and politicians (e.g., Gunia, 2019). The framing problem is compounded by the fact that a correction necessarily follows dissemination of a falsehood. This temporal sequencing is problematic in light of evidence that misinformation spreads faster and further online than true information (Vosoughi, Roy, & Aral, 2018). Corrections therefore inevitably play a catch-up game with misinformation and the corrections may be outpaced by falsehoods. Recent projections based on models of contemporary discourse on Facebook have raised the alarming possibility that anti-vaccination rhetoric may dominate the online landscape within a decade (Johnson et al., 2020).

It turns out that all these difficulties that beset even potentially successful debunking techniques can be circumvented by avoiding debunking altogether. Aside from *de*bunking, we should also explore *pre*bunking—that is, making people aware of potential misinformation *before* it is presented. This idea, known as inoculation, has a long history that has recently culminated in research that has yielded actionable knowledge for communicators.

Inoculation

Concern about people's general vulnerability to political indoctrination goes back many decades (McGuire, 1961), arising at the time from disquietude about persuasive techniques employed by totalitarian states. The larger question of how to go about developing attitudinal "resistance" against unwanted persuasion attempts ultimately led McGuire to develop "inoculation theory", which, for a popular audience, he described as a "vaccine for brainwash" (McGuire, 1970); see Figure 1.

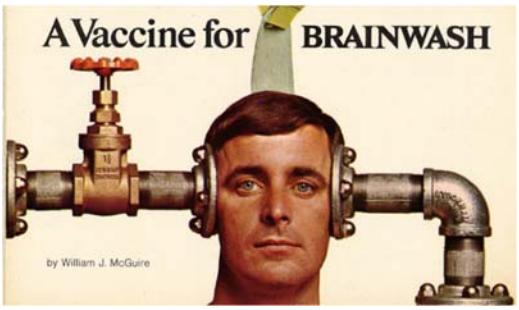

Figure 1. A Vaccine for Brainwash. From the original article by McGuire (1970) in *Psychology Today*. Copyright held by an unknown person.

Inoculation theory (Anderson & McGuire, 1965; McGuire, 1961, 1964, 1970; McGuire & Papageorgis, 1962) closely follows the biomedical analogy. Just as vaccines are weakened versions of a pathogen that trigger the production of antibodies when they enter the body to help confer immunity against future infection, inoculation theory postulates that the same can be achieved with information: by preemptively exposing people to a sufficiently weakened version of a persuasive attack, a cognitive-motivational process is triggered that is analogous to the production of "mental antibodies", rendering the individual more immune to persuasion (Compton, 2013; McGuire, 1961; Pfau, 1997).

Specifically, the psychological inoculation process consists of two core elements, including: 1) a warning to help activate threat in message recipients (to motivate resistance), and 2) refutational preemption (or *prebunking)*. These two components are assumed to work together in the following fashion: forewarning people that they are about to be exposed to challenging content is thought to elicit threat to motivate the protection of existing beliefs. In turn, two-sided refutational messages, which involve the threatening information, serve to both teach and inform people as they model the counterarguing process and provide specific content that can be used to resist persuasive attacks (Compton, 2013; McGuire, 1970). Over the last 50 years, a large body of

evidence across domains—from health to political campaigning—has revealed that inoculation messages can be effective at conferring resistance to persuasion. A meta-analysis by Banas and Rains (2010) that considered 40 studies with more than 10,000 participants altogether established an effect size of inoculation interventions of about $d = 0.43$ (conventionally considered to be "medium" in magnitude). Yet, although a handful of dedicated scholars have continued to publish on the theory (see Compton & Pfau, 2005; Pfau, 1997), interest among social psychologists has dwindled over the years.

As Eagly and Chaiken (1993) summarize in their landmark text on the psychology of attitudes, "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). Following Eagly and Chaiken's call that inoculation theory deserves renewed interest in the context of contemporary social psychological research, we outline our research program bringing inoculation theory into the 21st century. Importantly, although McGuire formulated his theories long before the rise of the internet, we now know that the propagation of misinformation through online social networks closely resembles the spread of a virus: rapidly transmitting highly infectious information from one host to another but without the need for physical contact (Budak, Agrawal, & El Abbadi, 2011; Kucharski, 2016). It must be of particular concern that false news on Twitter spreads faster, deeper, and broader than does truth (Vosoughi et al., 2018). Fake news appears to press several psychological hot buttons. One is negative emotions and how people express them online. For instance, Vosoughi et al. (2018) found that false stories were likely to inspire fear, disgust, and surprise; true stories, in contrast, triggered anticipation, sadness, joy, and trust. People are generally more likely to share messages featuring moral–emotional language (Brady, Wills, Jost, Tucker, & Van Bavel, 2017), and this tendency may be amplified by people's negativity bias, that is the human proclivity to attend more to negative than to positive things

(Soroka, Fournier, & Nir, 2019). The ability of false news to trigger negative emotions may thus give it an edge in the competition for human attention, and digital media may, as Crockett (2017) argued, promote the expression of negative emotions such as moral outrage "by inflating its triggering stimuli, reducing some of its costs and amplifying many of its personal benefits" (p. 769). Whether by design or coincidence, false online content appears to exploit these psychological factors.

The inoculation metaphor is therefore perhaps more relevant now than it was ever before, given that the natural antidote to a virus is the creation of a scalable vaccine. Accordingly, we outline three fundamental recent developments in inoculation theory scholarship that have pushed the theoretical boundaries of the original theory forward, namely; 1) a move away from a near-exclusive focus on "cultural truisms" towards inoculation against more contested issues, including fake news and misinformation, 2) a shift in focus from inoculation against specific arguments (narrow-spectrum) to the techniques that underlie manipulation and persuasion more generally (broad-spectrum), and 3) revisiting the potential of "active" vs. "passive" inoculation defenses. Our research program has enabled the vaccine metaphor to be scaled widely to address the real-world challenge of inoculating people against fake news and misinformation.

From cultural truisms to highly contested issues

There is a common (mis)perception that inoculation theory can only be applied to what McGuire (1970) referred to as "cultural truism" or "beliefs so generally accepted that most individuals are unaware of attacking arguments" (p. 37). Examples he gave included "the value of frequent tooth brushing" and "annual medical check-ups". Because student surveys indicated little polarization on these issues, uniformly favorable attitudes could therefore be strengthened against persuasive attacks through the process of inoculation. After all, if people had been

exposed to attitude dissonant information before on a topic, would this still constitute "preemptive" refutation? The overarching concern for McGuire was research on selective exposure: people tend to seek out information that will confirm their pre-existing view of the world and avoid information that conflicts with what they already believe. McGuire reasoned that if this is true, then people maintain their beliefs in what he called a "germ-free ideological" environment (i.e., they avoid contact with arguments that challenge their beliefs on controversial issues) and so inoculation would still apply. However, McGuire concluded that as a psychological mechanism, the literature on selective bias has a "questionable empirical status" (Anderson & McGuire, 1965, p. 46) as people do regularly seek out information that challenges their worldview and so it felt risky and premature to announce that inoculation would simply apply to all beliefs (McGuire, 1970).

Nonetheless, it is interesting that the focus of inoculation research—by and large—has remained with cultural truisms (Pfau et al., 2001), as this rigid interpretation of the initial metaphor hampers theory development. For example, consider that the threat element of the analogy has received intense debate, as it was unclear whether threat was meant to be elicited implicitly through exposure to a weakened attack (sending a warning signal to the mind, sort of speak, to help motivate antibody production) or whether it was meant to be implemented as an explicit forewarning. At any rate, McGuire initially did not test the threat component explicitly and it fits less clearly with the biological analogy (Compton, 2009). Yet, McGuire himself did actively encourage further pursuit of the medical analogy (McGuire & Papageorgis, 1962, p. 34). The consensus interpretation is therefore that the analogy is meant to be instructive rather than restrictive (Compton, 2019) to encourage further theoretical development and innovation. In fact, some 20 years after McGuire's initial experiments, Pryor and Steinfatt (1978) already noted that McGuire was incorrect about the fact that inoculation cannot be applied to issues where people

have differing prior beliefs, which has led a call to rethink the boundary conditions of the analogy more generally (Wood, 2007). Research by van der Linden et al. (2017) and Cook et al. (2017) addresses this directly. Both research teams showed that inoculation can be applied to one of the most contested issue in the United States today: global warming (Ballew, Goldberg, Rosenthal, Gustafson, & Leiserowitz, 2019).

Misinformation about climate change is rampant on the internet (e.g., Lewandowsky et al., 2019a). One potent online climate misinformation campaign is the "Global Warming Petition Project" (Cook, Maibach, van der Linden, & Lewandowsky, 2018). The petition engendered a viral misinformation story on social media in 2016 claiming that "tens of thousands of scientists have declared global warming a hoax" (Readfearn, 2016). In actual fact the petition was meaningless. The list contains no affiliations, making verification of signatories problematic (e.g., Charles Darwin and the Spice Girls were among the signatories; van der Linden et al., 2017). Fewer than 1% of the signatories have any expertise in climate science revealing the petition to be an instance of the "fake-experts" strategy that was pioneered by the tobacco industry in the 1970s and 1980s (Cook et al., 2017; Oreskes & Conway, 2010).

Although the petition has been debunked repeatedly, it continues to sow confusion. In a national probability sample of the United States population, van der Linden et al. (2017) found that amongst a wide range of fake claims, Americans were most persuaded by the debunked Oregon petition. Accordingly, in a subsequent experiment van der Linden et al. (2017) evaluated whether (a) such misinformation is actually harmful to public opinion formation and (b) if so, whether people can be inoculated against such specific falsehoods. In their online experiment (*N=2167*), participants were randomly assigned to one of five conditions. Figure 2 presents the data from the experiment and guides explanation of the conditions.

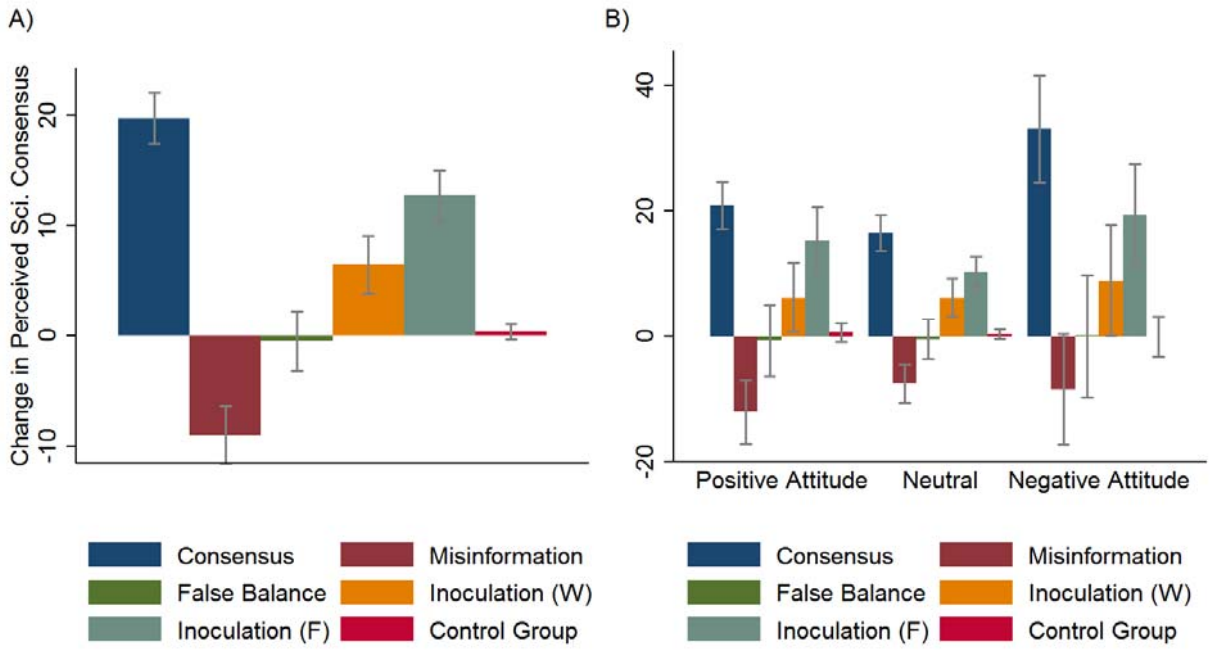

Figure 2. Inoculating against misinformation, adapted from van der Linden et al. (2017). Note: Error bars represent 95% confidence intervals. The three attitudinal groups were created based on answers to the pre-test questions, such that those who answered that they believe that climate change is happening *and* human-caused were classified as "positive", those who stated that they do not believe that climate change is happening at all were classified as "negative" and the remainder of the sample were classified as "neutral". The same patterns emerged for political party ID (Republican, Democrat, Independent).

The conditions were formed by presenting misinformation or factual information either alone or in combination. The factual information focused on the scientific consensus, namely the fact that over "97% of climate scientists have concluded that human-caused global warming is happening". Acceptance of that consensus had been identified by related research as a "gateway" for attitude change (Lewandowsky, Gignac, & Vaughan, 2013; van der Linden et al., 2019; van der Linden, 2015). The misinformation was a screenshot of the Global Warming Petition Project stating that "over 31,000 scientists have signed a petition that there is no scientific evidence for

human-caused global warming". In the experiment, participants were either exposed to just the scientific consensus (Figure 2A, "consensus"), just the misinformation by itself (Figure 2A, "misinformation"), a condition in which participants were first exposed to the scientific consensus before being exposed to the misinformation (Figure 2A, "false-balance") and two separate inoculation conditions. In the brief inoculation condition, participants were simply forewarned that politically motivated groups use misleading tactics to try to convince the public that there is a lot of disagreement between scientists (Figure 2A, "inoculation-W") whereas in the more detailed inoculation condition, the warning was accompanied by a traditional preemptive refutation of the petition by noting that many of the signatories are clearly fake (e.g. Charles Darwin), that although 31,000 may sounds big, it only comprises 0.3% of US science graduates, and that most of the signatories have no real expertise in climate science (Figure 2A, "inoculation-F").

The results showed that when participants were exposed to the full "dose" of the misinformation at the end of the experiment (i.e., the website of the petition), both inoculation conditions were effective in conferring attitudinal immunity against misinformation. In particular, although the misinformation itself proved potent—decreasing people's judgments about the scientific consensus in the absence of any inoculation ($d = 0.48$)—both the forewarning ($d = 0.33$) and full inoculation ($d = 0.75$) were effective in conferring resistance against the persuasive attack (maintaining about 1/3rd and 2/3^{rds} of the effect of the factual message). Although these results mainly speak the danger of misinformation and the efficacy of inoculation, strikingly, nearly the exact same patterns emerged regardless of people's prior attitude towards climate change (Figure 2B). In other words, the inoculation treatments equally protected against misinformation (and boosted belief in the scientific consensus) for those with positive, neutral, and negative prior attitudes toward the issue.

Although these results are not the first demonstration that inoculation works in the context of differing prior attitudes (e.g., see also Pryor & Steinfatt, 1978; Wood, 2007), or for an issue that is not a cultural truism (e.g., Banas & Miller, 2013; Jolley & Douglas, 2017), the highly politicized nature of the climate change debate pushes the boundary conditions of inoculation theory beyond what was previously thought possible.

There have been several additional recent extensions of the inoculation paradigm into contested arenas. For example, Zerback, Töpfl, and Knöpfle (2020) explored the effects of "astroturfed" comments launched by Russian "bots" on social media. Astroturfing refers to the manipulative use of media to create an artifial impression of grassroots support for an issue where no such support actually exists (McNutt & Boland, 2007). There is considerable evidence that Russian state-sponsored actors are engaged in astroturfing on social media (e.g., by amplifying public division in the context of vaccinations; Broniatowski et al., 2018). The primary astroturfing technique involves manufacturing of comments on social media that masquerade as authentic citizen voices. Zerback et al. (2020) showed in a large-scale experiment involving the German public that pro-Russian comments under a news article eroded people's belief that Russia was responsible for the Skripal poisoning in the U.K. This erosion of belief was preventable through inoculation, but only if the inoculation message anticipated the exact arguments to which participants were subsequently exposed—that is, the inoculation effect was specific rather than constituting a "broad spectrum" vaccine. Zerback et al. (2020) also showed that the inoculation effect wore off after a two-week delay (a similar wearing off of inoculation was reported by Niederdeppe, Gollust, & Barry, 2014).

Into the rabbit hole and beyond.

A particularly concerning manifestation of misinformation comprises conspiracy theories, which are often a gateway to extremism and radicalization. For example, the QAnon conspiracy theory, a contemporary instantiation of a "cabal theory" which holds that a single sinister group directs nearly all events in the world (Harari, 2020), has been identified as a security risk and domestic terror threat in the U.S (Amarasingam & Argentino, 2020). A conspiracy theory that links the 5G cellphone network to the emergence of COVID-19 has been associated with widespread vandalism of telecommunications installations in the U.K. in 2020. People who endorse this theory have been found to be willing to also endorse violence (Jolley & Paterson, 2020).

It is therefore encouraging that inoculation has been repeatedly found to be successful against conspiracy theories. Jolley and Douglas (2017) demonstrated the success of inoculation in an experiment involving people's attitudes towards vaccinations. In the inoculation condition, people were first exposed to anti-conspiratorial information which foreshadowed the arguments that conspiracy theorists might make against vaccinations, before being exposed to the conspiratorial material itself. In another condition, the order was reversed. Jolley and Douglas (2017) found that when people were inoculated by first receiving anti-conspiratorial material, they were no longer adversely affected by subsequent conspiratorial rhetoric. By contrast, if the conspiratorial material was presented first, the countering material was less effective. Similarly, Banas and Miller (2013) used both fact-based and logic-based inoculation material against a 9/11 conspiracy (the *Loose Change* film). Both approaches were found to be successful.

Inoculation has also been found to be successful against potential radicalization by online extremists. Braddock (2019) presented participants with pamphlets by rightwing and leftwing

extremist groups which, in the experimental conditions, were preceded by an inoculation treatment. The inoculation succeeded in making the extremist material unattractive in comparison to a no-treatment control condition. In a recent, as yet unpublished study by Muhsin Yesilada and the first author, inoculation was also found to be successful against Islamist and Islamophobic material. Participants who watched a brief training video that explained rhetorical techniques used by extremists were less likely to endorse subsequent radicalizing videos than people in the control condition who received no training. Similarly, in a recent study, Saleh et al. (2020) found that participants who were exposed to weakened doses of the strategies used in extremist recruitment—as part of the interactive inoculation game *Radicalize*—were more resistant and better able to identify manipulative social media messages when compared to a control condition.

In summary, recent research suggests that McGuire might have been surprised to learn that his initial reservations about the scope of inoculation theory were, in fact, conservative. There is now growing evidence that even controversial issues may be within the purview of the beneficial effects of inoculation. The shift toward contested issues has led scholars to rethink the original inoculation analogy by distinguishing between therapeutic and prophylactic inoculations (Compton, 2019). This distinction helped resolve a debate about whether inoculation in a contested domain still counts as “inoculation”, given that most people may have been exposed to arguments about climate change, online extremism, or high-profile events such as Russian responsibility for the poisoning of Sergei Skripal, a former Russian agent in the U.K. (Urban, 2019). In consequence, inoculation “in the wild” can hardly ever be truly preemptive (Basol, Roozenbeek, & van der Linden, 2020). From an outcome perspective this does not seem to matter much: Attitudes are protected from harmful information. From our perspective, the real-world inoculation process need not be inconsistent with the biomedical analogy. For example, consider that the incubation period for viral infections is highly variable, ranging from a couple of days up

to a few years, without a vaccine necessarily losing its effectiveness. The same could apply to how individuals become "infected" with misleading information. Moreover, developments of the psychological analogy can parallel those in medicine. Recent advances in medicine have found that therapeutic vaccines (which are administered after infection) can still reduce the effects of the disease by boosting immune response, for example in the context of HPV, hepatitis, and rabies (Autran, Carcelain, Combadiere, & Debre, 2004). As such, the distinction between prophylactic vs. therapeutic vaccines still allows for inoculation to occur within the context of differing prior attitudes and has opened up a completely new area of research (Compton, 2019).

From specific issues to broad-spectrum immunity.

One issue that has remained slightly unclear is the specificity of inoculation: Is it limited to the specific arguments that people might encounter later (Zerback et al., 2020), or might a cognitive "vaccine" provide "broad-spectrum" immunity; that is, might an inoculation message generalize to other arguments not previously encoutered? Recent research increasingly supports the latter alternative.

Using the exact same misinformation as van der Linden et al. (2017), Cook et al. (2017) conducted a similar inoculation study with national samples of the U.S. population ($N = 1{,}092$ and $N = 400$ in studies 1 and 2, respectively) with equally promising results. Cook et al. (2017) presented participants with (1) a warning that attempts are made to cast doubt on the scientific consensus on climate change for political reasons, and (2) an explanation that one disinformation technique involves appeals to dissenting "fake experts" to feign a lack of scientific consensus. Cook and colleagues illustrated the "fake-expert" approach by drawing attention to the historical attempts of the tobacco industry to undermine the medical consensus about the health risks from

smoking with advertising claims such as “20,679 Physicians say ‘Luckies are less irritating’”. Figure 3 shows the photo that accompanied the inoculation text in their experiment.

Figure 3. Stimulus used by Cook et al. (2017) to explain the disinformation strategy used by the tobacco industry to undermine the scientific consensus about the health risks from smoking. Reproduced from Cook et al. (2017) (Creative Commons, no permission required).

By exposing the fake-expert disinformation strategy at the outset, the subsequent misinformation (in this case, the feigned lack of scientific consensus on climate change) was defanged and people’s responses to various climate-related test items did not differ from a control condition that received no misinformation. By contrast, in the absence of inoculation, the misinformation involving “fake experts” had a discernible detrimental effect. An important further result of Cook et al. (2017) involves the role of political ideology, shown in Figure 4. On its own, misinformation had a polarizing effect such that Conservatives lowered their perception of the scientific consensus whereas Liberals’ perception remained unchanged (Figure 4, orange line). Because Liberals correctly estimated the consensus to be high, this implies that they were

unaffected by the misinformation whereas conservatives were susceptible to misinformation. There have been several recent reports that susceptibility to misinformation is greater on the populist right and among strong conservatives than the political left (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019; Andrew M. Guess et al., 2020; Guess, Nagler, & Tucker, 2019; Guess et al., 2020; Ognyanova, Lazer, Robertson, & Wilson, 2020; van der Linden, Panagopoulos, Azevedo, & Jost, 2020). The inoculation message administered before participants were exposed to the misinformation (Figure 4, blue line) completely neutralized its effect, thereby also eliminating the effect of participants' political ideology. This replicated the effect observed by van der Linden et al. (2017).

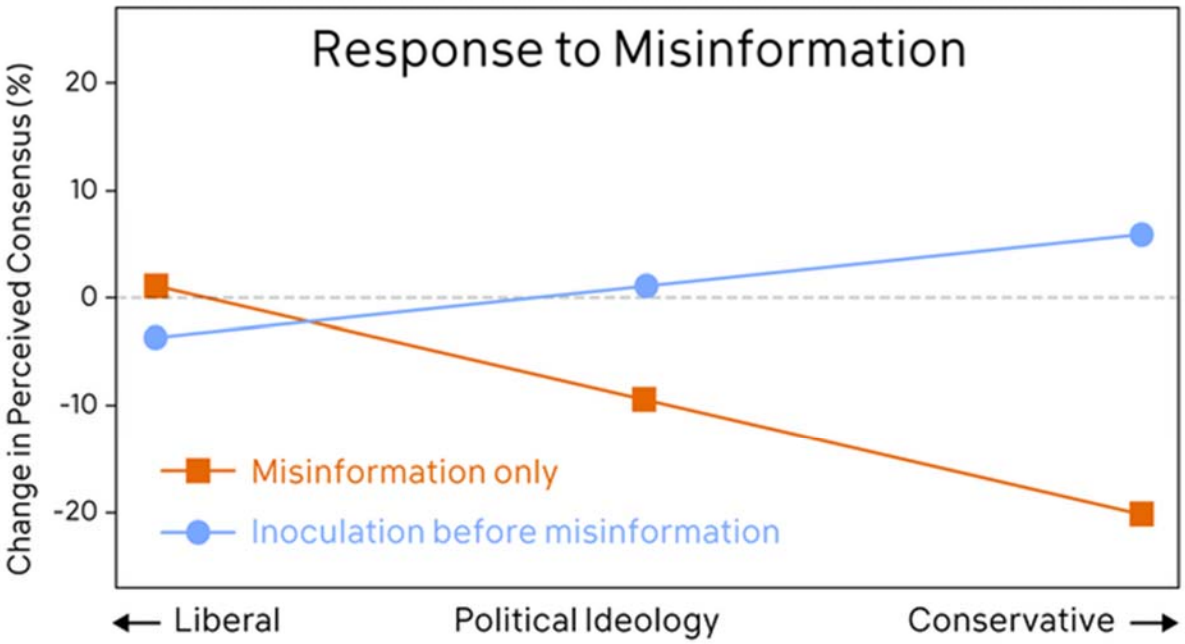

Figure 4. The effects of ideology on receptivity to misinformation (orange line) and its elimination by inoculation. Data were replotted by the authors from Cook et al. (2017). Note: political ideology was assessed with a measure of free-market support.

There is, however, an important difference between the procedures of van der Linden et al. (2017) and Cook et al. (2017). The procedure used by Cook and colleagues was not in the classical "refutational-same" format. In fact, their intervention did not mention the Global Warming Petition Project at all. Instead, their treatment inoculated participants by explaining a common manipulation technique: the promotion of fake experts. Cook et al. (2017) define the fake expert technique as "the use of spokespeople who convey the impression of expertise without possessing any relevant scientific expertise" (p. 11). This technique is not limited to the tobacco industry or climate denial. On the contrary, the technique is itself is widespread, for example, consider self-professed health experts advocating for homegrown cures against the coronavirus (such as gargling with lemon juice). The important result of Cook et al. (2017) is that exposing this technique in one context (medicine) inoculated individuals against the same technique in another context (climate change). This finding is crucial because it suggests the vaccine metaphor could be scaled by focusing less on specific issues and more on broader persuasion techniques. These findings accord with an emerging literature on "cross-protection" or the idea that an inoculation message can function as a "blanket of protection" by also conferring resistance to related yet untreated attitudes (Parker, Rains, & Ivanov, 2016). For example, Parker, Ivanov, and Compton (2012) showed that if young people were successfully inoculated against one health-adverse behavior (unprotected sex), the inoculation transferred to another risky behavior (binge drinking).

In the context of misinformation, it seems neither practical nor feasible to produce inoculations out of a weakened strain of a specific dose of fake news. Indeed, because fake news stories change and evolve on a frequent basis, this strategy would appear inefficient if applied at scale. In contrast, if a single inoculation treatment could offer widespread protection against a whole range of fake news, this would allow the analogy to be scaled and implemented more

easily. This notion of "generalized" resistance or a "broad-spectrum" vaccine was further developed in a series of studies in the first author's laboratory and by Roozenbeek and van der Linden (Roozenbeek & van der Linden Linden, 2018, 2019; Roozenbeek, van der Linden, & Nygren, 2020; van der Linden & Roozenbeek, 2020). Both lines of research suggest that rather than focusing on specific content, the public should be inoculated against the broader manipulation techniques that underlie the production of most misinformation. We turn to both lines of research in turn.

Inoculation by detecting flawed argumentation.

Researchers have compiled several inventories of flawed argumentation that are used to disinform, for example by populist politicians (Blassnig, Büchel, Ernst, & Engesser, 2019), anti-vaccination activists (Jacobson, Targonski, & Poland, 2007), or by people who spread conspiracy theories (Lewandowsky et al., 2015; Lewandowsky, Lloyd, & Brophy, 2018). The underlying rationale for those inventories is that, by and large, human cognition is a truth-tracking device. In many circumstances, cognition is found to be optimal by a Bayesian gold standard of rationality (e.g., Lewandowsky, Griffiths, & Kalish, 2009). Cognition that jettisons those normative standards is therefore likely to be less suited as a reality-tracking device, and its role in conspiracy theorizing and disinforming rhetoric is therefore unsurprising.

In the present context, it follows that if people can be trained to detect flawed argumentation, those skills might inoculate them against being misinformed in a fairly general "broad spectrum" manner. A stream of as-yet unpublished studies by the authors (in collaboration with Jon Roozenbeek and Google Jigsaw) has explored the use of brief (2-3 minute) videos to train people in the detection of flawed arguments. In all studies, participants in the inoculation condition were exposed to an argumentation-detection video that focused on a single misleading

technique, whereas in the control condition they watched a video about an unrelated issue (e.g., freezer burn). The template of each video consisted of both a forewarning as well as a weakened dose of the "virus" (i.e., a prebunk of the manipulation technique). In all studies, the inoculation improved participants' ability to detect misleading information, which in turn generally reduced their intention to share misleading material and increased discernment between trustworthy and untrustworthy material.

To illustrate, one of the techniques examined in our studies was incoherence. Incoherence is a frequent attribute of conspiracy theories (e.g., "Princess Diana was killed by MI5 and faked her own death"; Wood, Douglas, & Sutton, 2012) as well as climate denial (e.g., "Global temperatures cannot be measured accurately but we shouldn't worry because it has been cooling for the last 5 years"; Lewandowsky, Cook, & Lloyd, 2016). Incoherent arguments are, by definition, suspect and should be dismissed. Other techniques involved false dichotomies ("Either you are with us, or you are with the terrorists"; George W. Bush, 21 September 2001), scapegoating, ad hominem argumentation, and emotional manipulation.

In a slightly different context, Merpert, Furman, Anauati, Zommer, and Taylor (2018) showed that members of the public can be readily trained to identify statements in a politician's speech that could, in principle, be subject to fact checking. This is an important skill because opinions, by definition, are not subject to fact checking, and differentiation of opinions from factual assertions is therefore a necessary first step before fact-checking of suitable items can commence.

From passive to active inoculation: learning by doing.

McGuire initially hypothesized that compared to "passive" inoculation (where participants are simply provided with refutations to a particular argument), the inoculation process might be more effective when people are tasked with actively generating their own defenses or counter-arguments. In fact, inoculation messages are known to change the structure of associative memory networks, boosting nodes (e.g., counter-arguments) as well as the number of linkages between nodes, which helps strengthen people's ability to resist persuasion (Pfau et al., 2005).

Roozenbeek and van der Linden (2018, 2019) designed a real-world active inoculation simulation in the form of a free online "fake news game" called Bad News (www.getbadnews.com). The intervention simulates a social media feed and players are encouraged to step into the shoes of a fake news producer and over the course of 15 minutes gain as many followers as they can without losing credibility. The purpose of the game is to inoculate people against the techniques used in the production of fake news by letting them actively generate their own content in the simulation engine (see Figure 5). Roozenbeek and van der Linden identified 6 common manipulation techniques that are routinely involved in the production of fake news; impersonating people online (including experts), using emotional language (e.g., outrage), group polarization, floating conspiracy theories, discrediting opponents, and online trolling.

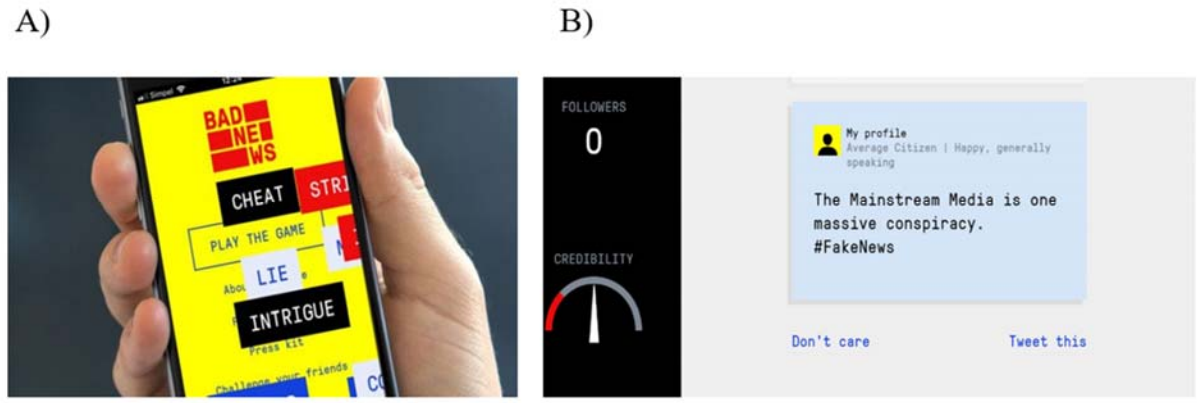

Figure 5. Screenshot of landing page (A) and gameplay (B). For further details visit www.getbadnews.com

The game shows players a meme or headline to which they can react in a number of ways. Progress in the game is measured through a "followers" and "credibility" meter (Figure 5B). Selecting an option that is consistent with what a "real" producer of disinformation would do earns players more followers and credibility. By contrast, if their strategy is too obvious or too much in line with journalistic best practice, the game either takes followers away or lowers players' credibility score. In the game, players start off by posting a tweet about something that frustrates them, which could be anything from the government, to the mainstream media, or the Flat Earth Society. Players then progress through 6 badges (or levels), each of which illustrates one of the manipulation techniques mentioned earlier; impersonation, emotion, polarization, conspiracy, discrediting, and trolling (for a detailed review of these techniques see Roozenbeek & van der Linden, 2019; van der Linden & Roozenbeek, 2020). The scenarios in which these techniques are defanged also make use of other popular concepts such as echo chambers and false amplification of a message. Players start the game by impersonating an official account, they can choose from various options such as impersonating Donald Trump (who declares war on North Korea) or NASA (which announces that a massive meteorite is about to hit earth). The game is

fully interactive and players are shown (simulated) reactions from other users and followers after they produce content. The game subsequently prompts the player to go professional and start their own news site by selecting a website name and slogan. The game was designed in collaboration with the Dutch media collective "DROG" and design studio Gusmanson. The game is based on full-cycle social psychology research (Mortensen & Cialdini, 2010), moving continuously from the lab to the field and back.

The game incorporates both elements of the inoculation process; (a) the game forewarns people that they are about to be exposed to challenging content and (b) the game exposes the player to severely weakened doses of the strategies that are used in the production of fake news. The doses are severely weakened through the use of ridicule and humor: they activate the immune system (getting the point across) but without actually overwhelming it (i.e., the content does not actually dupe people). The Bad News game has been played by about a million people worldwide (Roozenbeek et al., 2020b). The game features a research component where players are quizzed before and after gameplay on the reliability of fake and credible headlines using 7-point scales. Importantly, the test items are not featured in the game itself to help evaluate to what extent people can identify manipulation techniques in a range of "new" headlines. Although the test items are mirrored after real-world fake news, they are fictional for two important reasons: (1) to exclude memory and familiarity confounds (people may simply know a headline is real or fake because they have seen it before) and (2) to have sufficient experimental control over isolating and embedding the specific manipulation techniques in each of the test items.
An example item for the conspiracy badge asked players to judge reliability of the headline "The Bitcoin exchange rate is being manipulated by a small group of rich bankers" (which uses the conspiracy technique) or "New study shows that right-wing people lie far more than left-wing people" (which uses the polarization technique). An example of a credible real item that does not

make use of any of these techniques included; “Brexit, the United Kingdom’s exit from the European Union, will officially happen in 2019”. (The study was conducted before the U.K. formally exited the E.U., at a time when the exit date was thought to be in 2019. The U.K. ultimately departed on 31 January 2020.)

Roozenbeek and van der Linden (2019) initially evaluated the game using a within-subject design with a sample of roughly $N = 10{,}000$ people. The results are shown in Figure 6. For the real news items, people did not change their reliability ratings between a pre- and a post-test ($d = 0.03 - 0.04$, Figure 6D). For the fake news items, by contrast, people significantly downgraded reliability overall ($d = 0.52$) as well as for each technique separately (d ranges from 0.16 to 0.35, Figure 6A-C)[2]. Given that many elections are decided on small margins (e.g., half of U.S. presidential elections were decided by margins under 7.6% (Epstein & Robertson, 2015) and the 2016 election was decided by razor thin margins in a few swing states), these effects can be considered meaningful when scaled (Funder & Ozer, 2019) and commensurate with effect sizes in persuasion research (Banas & Rains, 2010; Walter & Murphy, 2018). Importantly, although Roozenbeek and van der Linden (2019) found some small variation in the inoculation effect across age and ideology, such that older people and Conservatives were slightly more susceptible to fake news on the pre-test (which is consistent with other recent work, e.g., Grinberg et al., 2019; Guess et al., 2019, 2020; for a review see Brashier & Schacter, 2020), the inoculation effect was significant across all subgroups.

---

[2] For a detailed methodological overview of item and testing effects using the *Bad News* paradigm we refer the reader to Roozenbeek et al. (2020b).

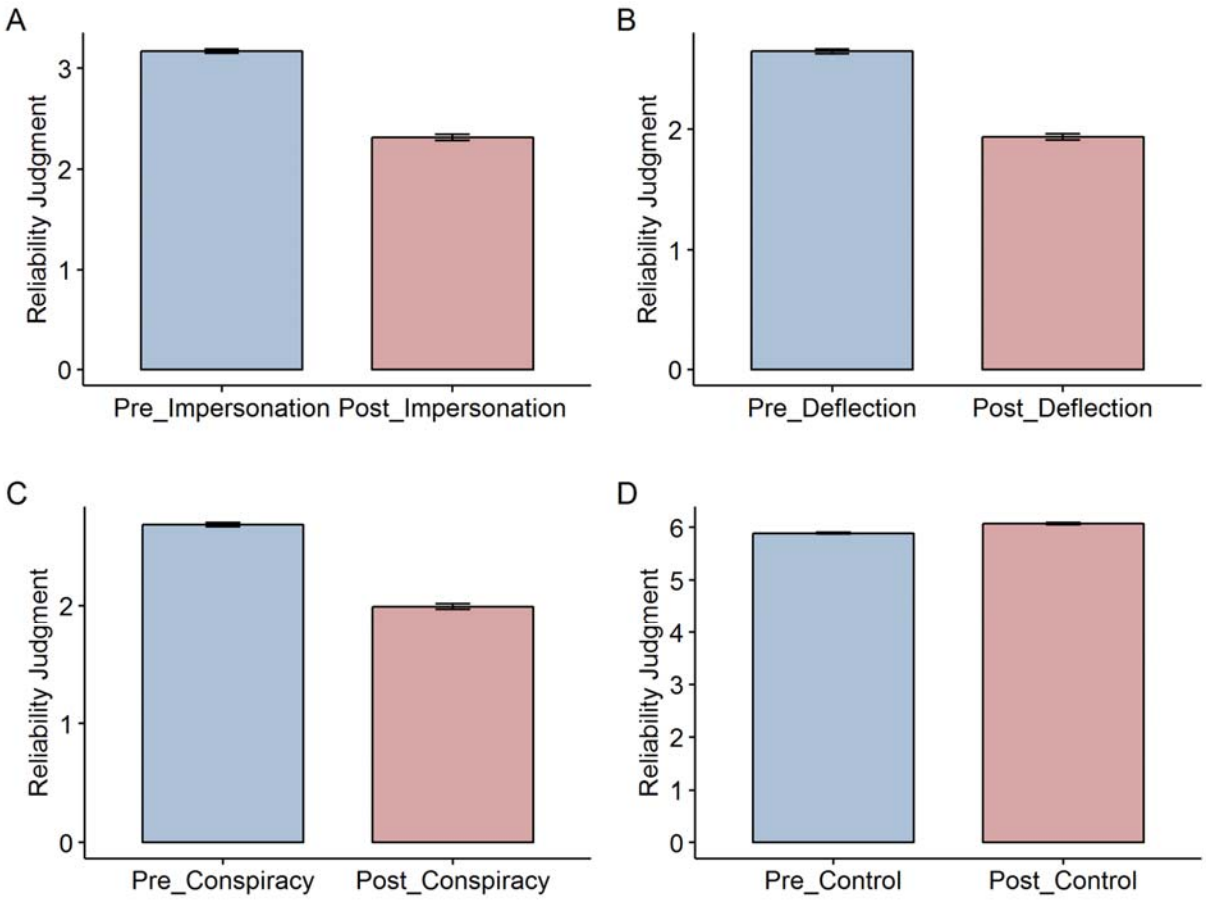

Figure 6. Pre and post scores for fake items that use manipulation techniques (panels A-C) as well as the mean score for the control items (panel D). Note: Error bars represent 95% confidence intervals. Adopted from Roozenbeek and van der Linden (2019).

Basol et al. (2020) replicated these findings in a randomized experiment with a treatment and a control group (the latter involved participants playing Tetris for 15 minutes). The results were very similar for the overall effect ($d = 0.60$) as well the range per technique ($d = 0.14$ to 0.45). Importantly, Basol et al. (2020) also included a measure of how much confidence players had in their judgments. Confidence plays a key role in the inoculation process (Tormala & Petty, 2004), as people who are confident in their beliefs are both more willing and able to defend them against persuasion attempts. Basol et al. (2020) found that the game significantly boosted

people's confidence in their judgments about the reliability of the fake items when those judgments were accurate ($d = 0.52$). Boosting of confidence is important because confidence in one's own beliefs is critical to being able to resist unwanted attempts to persuade and manipulate (Compton & Pfau, 2005).

The game has seen several spin-offs and real-world adaptations. For example, in collaboration with the UK government, the *Bad News* game has been translated worldwide into more than 15 languages to allow for larger-scale testing. Roozenbeek et al. (2020) were able to conduct a cross-cultural replication of the game in Sweden, Germany, Greece, and Poland. Although some cultural heterogeneity was observed, the principal effects of the intervention replicated overall across cultures. In 2020, Roozenbeek and van der Linden launched *GoViral!*, a game focused on prebunking COVID-19 misinformation specifically in collaboration with the UK Cabinet office with support from the WHO and UN (Reader, 2020), as well as *Harmony Square*, a game focused on inoculating against political misinformation during elections in collaboration with the Department of Homeland Security in the United States (Roozenbeek & van der Linden, 2020).

From a vaccine to herd immunity.

Many interesting questions remain, including how long the inoculation effect lasts. Inoculation treatments are typically observed to decay over a number of weeks (Banas & Rains, 2010; Niederdeppe et al., 2014; Zerback et al., 2020), much in line with the forgetting of conventional rebuttal efforts (Swire et al., 2017). Recent research has suggested that occasional booster doses can extend retention of inoculation (Ivanov, Parker, & Dillingham, 2018; Maertens, Anseel, & van der Linden, 2020). In the study by Maertens et al. (2020), the benefits of playing the "Bad News" game were found to wear off after 2 months without further interventions, but

the benefits retained intact for 3 months if the retention interval included a potential booster shot in the form of repeated testing.

Another open question is whether inoculation interacts with psychological reactance (though see Miller et al., 2013). Reactance refers to the motivational state that arises when people feel that their behavioral freedom has been threatened or taken away (Brehm & Brehm, 1981). When this occurs, individuals may act contrary to a prescribed action in order to protect or restore their feeling of freedom and control. It is unclear whether people who are high in trait reactance (e.g., Quick, Scott, & Ledbetter, 2011) are less receptive to inoculation messages. Attempts to inoculate against reactance (i.e., seeking to reduce the freedom threat of directive messages by inoculation) have been met with mixed success (Richards & Banas, 2015, 2015).

Although these questions open exciting and important avenues for future research, perhaps the most important question of all is how to translate a cognitive vaccine that boosts individual immune responses into societal level "herd immunity". Undoubtedly, the most powerful aspect of the inoculation metaphor was left relatively unexplored by McGuire; namely, the social nature of the theory (van der Linden, Maibach, Cook, Leiserowitz, & Lewandowsky, 2017). If enough individuals in a population are vaccinated, the informational virus has no opportunity to take hold and spread. Importantly, the metaphor implies that not every single individual needs to be vaccined, as herd immunity offers protection to those who are unable or unwilling to receive the vaccine. Accordingly, what is important about the newer (e.g., gamified) inoculation approaches is its ability to scale: the game can be shared interpersonally as well as on social media. In addition, the intervention is flexible and adaptive, and so scenarios can easily be changed and updated in response to new threats (e.g., deepfakes) to remain preemptive. In other words, just like misinformation, the vaccine can spread too, either because other people are enticed to play

the game or because people engage in something known as “post-inoculation talk”. Recent research has started to evaluate how interpersonal discussions following an inoculation intervention can strengthen attitude resistance through enhanced confidence and advocacy (Dillingham & Ivanov, 2016).

The potential for the social diffusion of inoculation content in social networks raises many exciting questions about how best to model its spread. For example, agent-based simulations are shedding light on how evidence-resistant minorties can delay consensus formation and undermine public opinion (Lewandowsky et al., 2019b). We expect that the future of inoculation theory scholarship will be best served by focusing on social psychological theories of how inoculation spreads from one person to another to be able to offer realistic predictions about the potential for attitudinal herd immunity against the increasing spread of fake news and misinformation.

Inoculating against manipulative personalization.

We conclude by turning attention to another arena of political communication that has been highly contested, namely “micro-targeting” of persuasive messages via Facebook or other social media. Micro-targeted political advertising exploits the unprecedented amounts of personal data that are harvested by platforms such as Facebook to reach its targets. There is evidence that knowledge of 300 Facebook “likes” is sufficient to infer a user’s personality with greater accuracy than their spouse (Youyou, Kosinski, & Stillwell, 2015). Micro-targeting erupted onto the public scence with the Cambridge Analytica scandal after the Brexit referendum in the U.K., when it transpired that the company had used profiles from 87 million Facebook users to target individuals with highly specific messages (Heawood, 2018). A British Parliamentary committee that investigated the scandal concluded that relentless targeting that plays “to the fears and the prejudices of people, in order to alter their voting plans” is “more invasive than obviously false

information" and contributes to a "democratic crisis" (Digital, Culture, Media and Sport Committee, 2019). Although Facebook has curtailed data access in response, advertisers can continue to select audiences on the basis of attributes that are now known to be predictive of personality. Content delivery can therefore continue to exploit, without users' awareness, sensitive details about their lives.

Although the impact of Cambridge Analytica is difficult to quantify, experimental evidence suggests that ads that are targeted at a person's personality are more effective than other ads. In a large-scale "real life" experiment on Facebook, Matz, Kosinski, Nave, and Stillwell (2017) showed that cosmetic ads that were designed to appeal either to introverts or to extraverts (Figure 7A, top and bottom, respectively) elicited more click-throughs and purchases when recipients' personality matched than when it did not.[3]

[3] The study by Matz et al. (2017) has been subjected to critiques (Eckles, Gordon, & Johnson, 2018; Sharp, Danenberg, & Bellman, 2018) which were (in our view) convincingly rebutted by the authors (Matz et al., 2018a, 2018b).

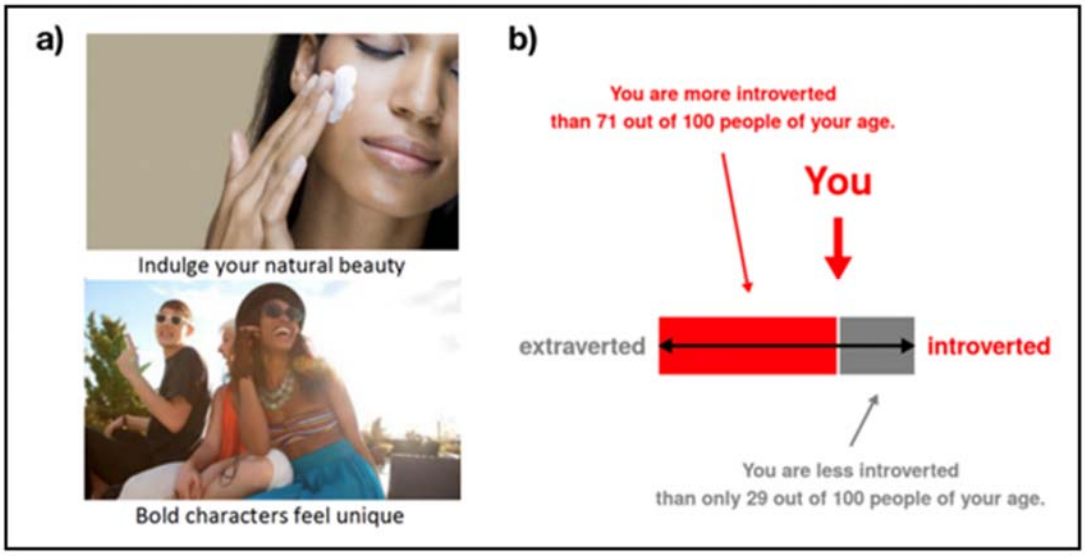

Figure 7. (A) Advertisements designed by Matz et al. (2017) that target introverts (top) and extraverts. (B) Feeback provided to participants in the experiment by Lorenz-Spreen and colleagues.

Although advertisements for cosmetics are unlikely to alter the course of history, they nonetheless open a window into the power of algorithmic targeting on social media. It is therefore important to ask whether people might be protected against targeted manipulation by "boosting" their detection skills: might the provision of information about their personality inoculate a person against inadvertently being particularly receptive to a targeted ad? An as-yet unpublished experiment involving the first author (Lorenz-Spreen, Hertwig, Lewandowsky, & Herzog, in preparation) showed that this is indeed possible. In the experimental "boosting" condition, participants were provided with information about their introversion-extraversion score (Figure 7B) together with a brief explanation of the characteristics of the two personality types. During a subsequent classification task, in which participants had to decide for each ad whether or not it matched their personality, performance was considerably better in the boosting condition than in a control condition involving feedback about an unrelated personal attribute.

Conclusion

We live in an environment that is drenched in misinformation, "fake news", and propaganda not because of an unavoidable accident but because it has been created by political actors in pursuit of political and economic objectives (Lewandowsky, 2020; Lewandowsky et al., 2017b). We therefore do not face a natural disaster but a political problem. On the positive side, this implies that, unlike for earthquakes or tsunamis, a solution is likely to exist and ought to be achievable. On the negative side, it means that the solution is unlikely to involve more (or better) communication alone. As Brulle, Carmichael, and Jenkins (2012) noted in the context of climate change, "introducing new messages or information into an otherwise unchanged socioeconomic system will accomplish little" (p. 185). Instead, we need to pursue multiple avenues—many of them political—to contain misinformation and redesign the information architecture that facilitate its dissemination (Kozyreva, Lewandowsky, & Hertwig, 2020; Lorenz-Spreen, Lewandowsky, Sunstein, & Hertwig, 2020). van der Linden (2019) postulated several such behavioral avenues, starting with prebunking or inoculation, which is followed where necessary by real-time rebuttal or fact-checking and then debunking if inoculation fails. Lorenz-Spreen et al. (2020) additionally provided an analysis of how online architectures contribute to the spread of misinformation and how they could be redesigned to facilitate accurate democratic deliberation. In short, future work would be well-served by adopting a multi-layered response to misinformation, including the techniques that we have reviewed here.

It is encouraging that inoculation techniques have been successful in the "real world" outside the laboratory. For example, during a mumps epidemic in Iowa in 2006, the Department of Public Health posted a primer for the media online. The primer provided explanations and rebuttals to anticipated arguments by anti-vaccine activists. This enabled the media to understand

and defang those contrarian arguments (Jacobson et al., 2007). We invite psychologists of all stripes to consider the benefits of inoculation in eradicating the spread of misinformation.

References

Amarasingam, A., & Argentino, M.-A. (2020). The QAnon conspiracy theory: A security threat in the making? *CTC Sentinel*, *13*(7), 37–44.

Anderson, L. R., & McGuire, W. J. (1965). Prior reassurance of group consensus as a factor in producing resistance to persuasion. *Sociometry*, 44–56.

Arenge, A., Lapinski, J., & Tallevi, A. (2018, May). Poll: Republicans who think Trump is untruthful still approve of him. *NBC News*. https://www.nbcnews.com/politics/politics-news/poll-republicans-who-think-trump-untruthful-still-approve-him-n870521.

Autran, B., Carcelain, G., Combadiere, B., & Debre, P. (2004). Therapeutic vaccines for chronic infections. *Science*, *305*, 205–208. doi:10.1126/science.1100600

Babalola, S., Krenn, S., Rimal, R., Serlemitsos, E., Shaivitz, M., Shattuck, D., & Storey, D. (2020). KAP COVID dashboard. Johns Hopkins Center for Communication Programs, Massachusetts Institute of Technology, Global Outbreak Alert; Response Network, Facebook Data for Good. Retrieved from https://ccp.jhu.edu/kap-covid/kap-covid-trend-analysis-for-23-countries/

Ballew, M. T., Goldberg, M. H., Rosenthal, S. A., Gustafson, A., & Leiserowitz, A. (2019). Systems thinking as a pathway to global warming beliefs and attitudes through an ecological worldview. *Proceedings of the National Academy of Sciences*, *116*, 8214–8219. doi:10.1073/pnas.1819310116

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, *39*, 184–207. doi:10.1111/hcre.12000

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*, 281–311.

Basol, M., Roozenbeek, J., & Linden, S. V. der. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*, 1–9. doi:10.5334/joc.91

Blassnig, S., Büchel, F., Ernst, N., & Engesser, S. (2019). Populism and informal fallacies: An analysis of right-wing populist rhetoric in election campaigns. *Argumentation*, *33*, 107–136. doi:10.1007/s10503-018-9461-2

Braddock, K. (2019). Vaccinating against hate: Using attitudinal inoculation to confer resistance to persuasion by extremist propaganda. *Terrorism and Political Violence*. doi:10.1080/09546553.2019.1693370

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*, 7313–7318. doi:10.1073/pnas.1618923114

Brashier, N. M., & Schacter, D. L. (2020). Aging in a fake news era. *Current Directions in Psychological Science*.

Brehm, S. S., & Brehm, J. W. (1981). *Psychological reactance: A theory of freedom and control.* New York: Academic Press.

Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation*. Reuters Institute, University of Oxford.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., … Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, e1–e7. doi:10.2105/AJPH.2018.304567

Brulle, R. J., Carmichael, J., & Jenkins, J. C. (2012). Shifting public opinion on climate change: An empirical assessment of factors influencing concern over climate change in the U.S., 2002–2010. *Climatic Change*, *114*, 169–188. doi:0.1007/s10584-012-0403-y

Budak, C., Agrawal, D., & El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on world wide web - WWW '11*. doi:10.1145/1963405.1963499

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*. doi:10.1177/0956797617714579

Compton, J. (2009). Threat explication: What we know and don't yet know about a key component of inoculation theory. *STAM Journal*, *39*, 1–18.

Compton, J. (2013). Inoculation theory. In J. Dillard & L. Shen (Eds.), *The SAGE handbook of persuasion: Developments in theory and practice*. SAGE Publications, Inc. doi:10.4135/9781452218410

Compton, J. (2019). Prophylactic versus therapeutic inoculation treatments for resistance to influence. *Communication Theory*. doi:10.1093/ct/qtz004

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*, 97–146. doi:10.1080/23808985.2005.11679045

Compton, J., van der Linden, S., Cook, J., & Basol, M. (2019). Inoculation theory and science communication: Extant findings and new directions. Presented at the Paper presented at the 69th conference of the International Communication Association, Washington, D.C.

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

Cook, J., Maibach, E., van der Linden, S., & Lewandowsky, S. (2018). The consensus handbook. doi:10.13021/G8MM6P

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769–771. doi:10.1038/s41562-017-0213-3

DeStefano, F., & Thompson, W. W. (2004). MMR vaccine and autism: An update of the scientific evidence. *Expert Review of Vaccines*, *3*, 19–22. doi:10.1586/14760584.3.1.19

Digital, Culture, Media and Sport Committee. (2019). *Disinformation and "fake news": Final report*. House of Commons, U.K. Parliament. Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179102.htm

Dillingham, L. L., & Ivanov, B. (2016). Using postinoculation talk to strengthen generated resistance. *Communication Research Reports*, *33*, 295–302. doi:10.1080/08824096.2016.1224161

Dixit, P., & Mac, R. (2018). How WhatsApp destroyed a village. BuzzFeed News. Retrieved from https://www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich.

Ecker, U. K. H., Lewandowsky, S., Cheung, C. S. C., & Maybery, M. T. (2015). He did it! She did it! No, she did not! Multiple causal explanations and the continued influence of misinformation. *Journal of Memory and Language*, *85*, 101–115.

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

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*, 1087–1100. doi:10.3758/MC.38.8.1087

Eckles, D., Gordon, B. R., & Johnson, G. A. (2018). Field studies of psychologically targeted ads face threats to internal validity. *Proceedings of the National Academy of Sciences*, *115*, E5254–E5255. doi:10.1073/pnas.1805363115

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, *112*, E4512–E4521. doi:10.1073/pnas.1419828112

Fein, S., McCloskey, A. L., & Tomlinson, T. M. (1997). Can the jury disregard that information? The use of suspicion to reduce the prejudicial effects of pretrial publicity and inadmissible testimony. *Personality and Social Psychology Bulletin*, *23*, 1215–1226. doi:10.1177/01461672972311008

Freeman, D., Loe, B. S., Chadwick, A., Vaccari, C., Waite, F., Rosebrock, L., … al. (2020). COVID-19 vaccine hesitancy in the UK: The Oxford coronavirus explanations, attitudes, and narratives survey (OCEANS) II. *Psychological Medicine*, 1–34. doi:10.1017/S0033291720005188

Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., … Lambe, S. (2020). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological Medicine*. doi:10.1017/s0033291720001890

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*, 156–168. doi:10.1177/2515245919847202

Gills, B., & Morgan, J. (2020). Global Climate Emergency: After COP24, climate science, urgency, and the threat to humanity. *Globalizations*, *17*, 885–902. doi:10.1080/14747731.2019.1669915

Godlee, F., Smith, J., & Marcovitch, H. (2011). Wakefield's article linking MMR vaccine and autism was fraudulent: Clear evidence of falsification of data should now close the door on this damaging vaccine scare. *BMJ: British Medical Journal*, *342*, 64–66.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. Presidential election. *Science*, *363*, 374–378. doi:10.1126/science.aau2706

Guess, A., & Coppock, A. (2018). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*. doi:10.1017/s0007123418000327

Guess, A. M., Lockett, D., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2020). "Fake news" may have limited effects on political participation beyond increasing beliefs in false claims. *Harvard Kennedy School Misinformation Review*. doi:10.37016/mr-2020-004

Guess, A. M., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*, eaau4586. doi:10.1126/sciadv.aau4586

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 U.S. election. *Nature Human Behavior*, *in press*.

Gunia, A. (2019). The U.K. has officially declared a climate "emergency". TIME. Retrieved from https://time.com/5581615/uk-declares-climate-emergency/

Harari, Y. N. (2020). When the world seems like one big conspiracy. New York Times. Retrieved from https://www.nytimes.com/2020/11/20/opinion/sunday/global-cabal-conspiracy-theories.html

Heawood, J. (2018). Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity*, *23*, 429–434. doi:10.3233/IP-180009

Ivanov, B., Parker, K. A., & Dillingham, L. L. (2018). Testing the limits of inoculation-generated resistance. *Western Journal of Communication*, *82*, 648–665. doi:10.1080/10570314.2018.1454600

Jacobson, G. C. (2010). Perception, memory, and partisan polarization on the iraq war. *Political Science Quarterly*, *125*, 31–56.

Jacobson, R. A., Targonski, P. V., & Poland, G. A. (2007). A taxonomy of reasoning flaws in the anti-vaccine movement. *Vaccine*, *25*, 3146–3152. doi:10.1016/j.vaccine.2007.01.046

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1420–1436.

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., Oud, S. E., … Lupu, Y. (2020). The online competition between pro- and anti-vaccination views. *Nature*. doi:10.1038/s41586-020-2281-1

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*, 459–469. doi:10.1111/jasp.12453

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*. doi:10.1111/bjso.12394

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest*, *21*, 103–156. doi:10.1177/1529100620946707

Kucharski, A. (2016). Study epidemiology of fake news. *Nature*, *540*, 525–525. doi:10.1038/540525a

Kull, C., S. And Ramsay, Stephens, A., Weber, S., Lewis, E., & Hadfield, J. (2006). Americans on Iraq: Three years on. *Program on International Policy Attitudes*, *0*, 1–19.

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L., & Ratzan, S. C. (2011). Addressing the vaccine confidence gap. *The Lancet*, *378*, 526–535.

Lewandowsky, S. (2020). Wilful construction of ignorance: A tale of two ontologies. In R. Hertwig & C. Engel (Eds.), *Deliberate ignorance: Choosing not to know* (pp. 101–117). Cambridge, MA: MIT Press.

Lewandowsky, S., & Cook, J. (2020). Coronavirus conspiracy theories are dangerous—here's how to stop them spreading. The Conversation. Retrieved from https://theconversation.com/coronavirus-conspiracy-theories-are-dangerous-heres-how-to-stop-them-spreading-136564

Lewandowsky, S., Cook, J., & Ecker, U. K. H. (2017a). Letting the gorilla emerge from the mist: Getting past post-truth. *Journal of Applied Research in Memory and Cognition*, *6*, 418–424. doi:10.1016/j.jarmac.2017.11.002

Lewandowsky, S., Cook, J., Fay, N., & Gignac, G. E. (2019a). Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Memory & Cognition*, *47*, 1445–1456. doi:10.3758/s13421-019-00948-y

Lewandowsky, S., Cook, J., & Lloyd, E. (2016). The "Alice in Wonderland" mechanics of the rejection of (climate) science: Simulating coherence by conspiracism. *Synthese*, *195*, 175–196. doi:10.1007/s11229-016-1198-6

Lewandowsky, S., Cook, J., Oberauer, K., Brophy, S., Lloyd, E. A., & Marriott, M. (2015). Recurrent fury: Conspiratorial discourse in the blogosphere triggered by research on the role of conspiracist ideation in climate denial. *Journal of Social and Political Psychology*, *3*, 142–178. doi:10.5964/jspp.v3i1.443

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

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

Lewandowsky, S., Gignac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*, 399–404. doi:10.1038/nclimate1720

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*, 969–998. doi:10.1111/j.1551-6709.2009.01045.x

Lewandowsky, S., Lloyd, E. A., & Brophy, S. (2018). When THUNCing Trumps thinking: What distant alternative worlds can tell us about the real world. *Argumenta*, *3*, 217–231. doi:10.23811/52.arg2017.lew.llo.bro

Lewandowsky, S., Pilditch, T. D., Madsen, J. K., Oreskes, N., & Risbey, J. S. (2019b). Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. *Cognition*, *188*, 124–139. doi:10.1016/j.cognition.2019.01.011

Lewandowsky, S., Stritzke, W. G. K., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. *Psychological Science*, *16*, 190–195. doi:10.1111/j.0956-7976.2005.00802.x

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth and, autonomy and democratic discourse online. *Nature Human Behaviour*, *4*, 1102–1109. doi:10.1038/s41562-020-0889-7

MacLeod, C., Winter, M., & Gray, A. (2014). Beijing-bound flight from Malaysia missing. USA Today. Retrieved from https://eu.usatoday.com/story/news/world/2014/03/07/malaysia-airlines-beijing-flight-missing/6187779/

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*, 101455. doi:10.1016/j.jenvp.2020.101455

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.* Advance online publication. doi: 10.1037/xap0000315.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *48*, 12714–12719. doi:10.1073/pnas.1710966114

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2018a). Reply to Sharp et al.: Psychological targeting produces robust effects. *Proceedings of the National Academy of Sciences*, *115*, E7891–E7891. doi:10.1073/pnas.1811106115

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2018b). Reply to Eckles et al.: facebook's optimization algorithms are highly unlikely to explain the effects of psychological targeting. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1806854115

McCright, A. M., Charters, M., Dentzman, K., & Dietz, T. (2016). Examining the effectiveness of climate change frames in the face of a climate change denial counter-frame. *Topics in Cognitive Science*, *8*, 76–97. doi:10.1111/tops.12171

McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *The Journal of Abnormal and Social Psychology*, *63*, 326–332. doi:10.1037/h0048344

McGuire, W. J. (1964). Some contemporary approaches. In *Advances in experimental social psychology* (pp. 191–229). Elsevier. doi:10.1016/s0065-2601(08)60052-0

McGuire, W. J. (1970). Vaccine for brainwash. *Psychology Today*, *3*(9), 36–64.

McGuire, W. J., & Papageorgis, D. (1962). Effectiveness of forewarning in developing resistance to persuasion. *Public Opinion Quarterly*, *26*, 24–34. doi:10.1086/267068

McNutt, J., & Boland, K. (2007). Astroturf, technology and the future of community mobilization: Implications for nonprofit theory. *The Journal of Sociology & Social Welfare*, *34*, 165–178.

Merpert, A., Furman, M., Anauati, M. V., Zommer, L., & Taylor, I. (2018). Is that even checkable? An experimental study in identifying checkable statements in political discourse. *Communication Research Reports*, *35*, 48–57. doi:10.1080/08824096.2017.1366303

Miller, C. H., Ivanov, B., Sims, J., Compton, J., Harrison, K. J., Parker, K. A., … Averbeck, J. M. (2013). Boosting the potency of resistance: Combining the motivational forces of inoculation and psychological reactance. *Human Communication Research*, *39*, 127–155.

Mortensen, C. R., & Cialdini, R. B. (2010). Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass*, *4*, 53–63. doi:10.1111/j.1751-9004.2009.00239.x

Mozur, P. (2018). A genocide incited on Facebook, with posts from Myanmar's military. The New York Times. Retrieved from https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Murphy, G., Loftus, E. F., Grady, R. H., Levine, L. J., & Greene, C. M. (2019). False memories for fake news during Ireland's abortion referendum. *Psychological Science*. doi:10.1177/0956797619864887

Niederdeppe, J., Gollust, S. E., & Barry, C. L. (2014). Inoculation in competitive framing examining message effects on policy preferences. *Public Opinion Quarterly*, *78*, 634–655. doi:10.1093/poq/nfu026

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*. doi:10.1007/s11109-019-09528-x

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330.

OED. (2016). Retrieved from https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016)

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*. doi:10.37016/mr-2020-024

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt*. London, UK: Bloomsbury Publishing.

Parker, K. A., Ivanov, B., & Compton, J. (2012). Inoculation's efficacy with young adults' risky behaviors: Can inoculation confer cross-protection over related but untreated issues? *Health Communication*, *27*, 223–233. doi:10.1080/10410236.2011.575541

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*, 49–68. doi:10.1080/03637751.2015.1030681

Peretti-Watel, P., Seror, V., Cortaredona, S., Launay, O., Raude, J., Verger, P., … Ward, J. K. (2020). A future vaccination campaign against COVID-19 at risk of vaccine hesitancy and politicisation. *The Lancet Infectious Diseases*. doi:10.1016/s1473-3099(20)30426-6

Pfau, M. (1997). The inoculation model of resistance to influence. *Progress in communication sciences*, *13*, 133–172.

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*, 414–441. doi:10.1080/03637750500322578

Pfau, M., Szabo, A., Anderson, J., Morrill, J., Zubric, J., & Wan, H. H. (2001). The role and impact of affect in the process of resistance to persuasion. *Human Communication Research*, *27*, 216–252.

Poland, G. A., & Spier, R. (2010). Fear, misinformation, and innumerates: How the Wakefield paper, the press, and advocacy groups damaged the public health. *Vaccine*, *28*, 2361–2362.

Pryor, B., & Steinfatt, T. M. (1978). THE EFFECTS OF INITIAL BELIEF LEVEL ON INOCULATION THEORY AND ITS PROPOSED MECHANISMS. *Human Communication Research*, *4*, 217–230. doi:10.1111/j.1468-2958.1978.tb00611.x

Quick, B. L., Scott, A. M., & Ledbetter, A. M. (2011). A close examination of trait reactance and issue involvement as moderators of psychological reactance theory. *Journal of Health Communication*, *16*, 660–679. doi:10.1080/10810730.2011.551989

Quinnipiac. (2018). U.S. voters give Trump highest grade ever on economy. Quinnipiac University. Retrieved from https://poll.qu.edu/national/release-detail?ReleaseID=2587

Reader, R. (2020, October 13th). This game can stop people from falling for COVID-19 conspiracies. *FastCompany.* Retrieved from https://www.fastcompany.com/90563255/covid-19-conspiracies-go-viral-game

Readfearn, G. (2016). Revealed: Most popular climate story on social media told half a million people the science was a hoax. DeSmogBlog. Retrieved from https://www.desmogblog.com/2016/11/29/revealed-most-popular-climate-story-social-media-told-half-million-people-science-was-hoax

Richards, A. S., & Banas, J. A. (2015). Inoculating against reactance to persuasive health messages. *Health Communication*, *30*, 451–460. doi:10.1080/10410236.2013.867005

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. doi:10.1080/13669877.2018.1443491

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Nature Humanities and Social Sciences Communications*, *5* (65). doi: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 Misinformation Review 1*(8). doi: 10.37016/mr-2020-47

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review 1(2)*. doi:10.37016//mr-2020-008

Roozenbeek, J., Maertens, R., McClanahan, W.P., & van der Linden, S. (2020a). Differentiating item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educational and Psychological Measurement.*

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., ... & van der Linden, S. (2020b). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10), 201199. doi: 10.1098/rsos.201199

Schaeffer, K. (2020). Nearly three-in-ten Americans believe COVID-19 was made in a lab. Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2020/04/08/nearly-three-in-ten-americans-believe-covid-19-was-made-in-a-lab/

Sharp, B., Danenberg, N., & Bellman, S. (2018). Psychological targeting. *Proceedings of the National Academy of Sciences*, *115*, E7890–E7890. doi:10.1073/pnas.1810436115

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*, 18888–18892. doi:10.1073/pnas.1908369116

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science*, *4*, 160802. doi:10.1098/rsos.160802

Swire, B., & Ecker, U. K. H. (2017). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In B. Southwell, E. A. Thorson, & L. Sheble (Eds.), *Misinformation and mass audiences. Austin, tx: University of texas press.* Austin, TX: University of Texas Press.

Swire-Thompson, B., Ecker, U. K. H., Lewandowsky, S., & Berinsky, A. J. (2020). They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political Psychology*, *41*, 21–34. doi:10.1111/pops.12586

Tormala, Z. L., & Petty, R. E. (2004). Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, *14*, 427–442. doi:10.1207/s15327663jcp1404_11

Urban, M. (2019). Skripal poisoning: Third Russian suspect "commanded attack". BBC. Retrieved from https://www.bbc.co.uk/news/uk-48801205

van der Linden, S. (2019). Countering science denial. *Nature Human Behaviour*, *3*(9), 889-890. doi: 10.1038/s41562-019-0631-5

van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PloS One*, *10*(2), e0118489. doi:10.1371/journal.pone.0118489

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1(2)*, 1600008. doi:10.1002/gch2.201600008

van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. *Journal of Environmental Psychology*, *62*, 49–58. doi:10.1016/j.jenvp.2019.01.009

van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., & Lewandowsky, S. (2017). Inoculating against misinformation. *Science*, *358* (6367), 1141–1142. doi:10.1126/science.aar4533

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: Political bias in perceptions of fake news. *Media, Culture & Society 42*(3), 460-470. doi:10.1177/0163443720906992

van der Linden, S., & Roozenbeek, J. (2020). Psychological inoculation against fake news. In R. Greifeneder, M. Jaffé, E.J. Newman, & N. Schwarz (Eds.), *The psychology of fake news:*

Accepting, sharing, and correcting misinformation (pp 147-169). London, UK: Psychology Press.

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, *11*, 2928. doi: 10.3389/fpsyg.2020.566790

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151. doi:10.1126/science.aap9559

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

Watson, L. (2018). Systematic epistemic rights violations in the media: A Brexit case study. *Social Epistemology*, *32*, 88–102. doi:10.1080/02691728.2018.1440022

Wilkes, A. L., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *40*, 361–387.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social Psychological and Personality Science*, *3*, 767–773. doi:10.1177/1948550611434786

Wood, M. L. M. (2007). Rethinking the inoculation analogy: Effects on subjects with differing preexisting attitudes. *Human Communication Research*, *33*, 357–378. doi:10.1111/j.1468-2958.2007.00303.x

Wood, T., & Porter, E. (2018). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*. doi:10.1007/s11109-018-9443-y

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*, 1036–1040. doi:10.1073/pnas.1418680112

Zerback, T., Töpfl, F., & Knöpfle, M. (2020). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society*. doi:10.1177/1461444820908530