ISBN: 978-9934-564-49-9

INOCULATION THEORY AND MISINFORMATION

Published by the
NATO Strategic Communications
Centre of Excellence

ONLINE ISBN: 978-9934-564-49-9
PRINT ISBN: 978-9934-564-44-4
Authors: Dr Jon Roozenbeek and Professor Sander van der Linden,
Cambridge Social Decision-Making Lab, Department of Psychology, University of Cambridge
Project manager: Henrik Twetman
Copy Editing: Tomass Pildegovičs
Design: Linda Curika

Riga, October 2021
NATO STRATCOM COE
11b Kalnciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook/stratcomcoe
Twitter: @stratcomcoe

Author's Acknowledgements. We are grateful for research funding from Jigsaw, the IRIS Coalition (UK Government, #SCH-00001-3391), the Economic and Social Research Council (ESRC, #ES/V011960/1), and JITSUVAX (EU 2020 Horizon #964728). We are also grateful to DROG and Gusmanson for their key role in the design and development of the Bad News, Harmony Square, and Go Viral! Games. We also wish to thank our research collaborators. Finally, the research reported here was conducted in no small part by current and former members of the Cambridge Social Decision-Making Lab, and we are grateful for all their work.

INTRODUCTION

The spread of false and misleading information both online and offline poses a threat to the well-being of individuals, democratic institutions, and societies around the world (1, 2). The harmful consequences of the spread of false and/or misleading information can be seen in the proliferation of anti-vax groups on Facebook (3, 4), lack of confidence in the science of climate change (5), acts of vandalism committed on the basis of false conspiracy theories about COVID-19 (6, 7), and its influence on the exacerbation of radicalisation and polarisation (8, 9). However, a lack of consensus with respect to what constitutes "misinformation" (for example: focusing exclusively on false information is problematic because truth value can be difficult to determine objectively, and misleading and/or hyperpartisan content may significantly outweigh "fake news", see 10) and disagreement about the efficacy of various efforts to mitigate the spread of misinformation through algorithms, legislation, or content moderation means that no single intervention is likely to be sufficient (11, 12). In this report, we explore the role that psychology and behavioural science can play in the mitigation of online misinformation. To do so, we first discuss how to define "misinformation", and how it relates to various other commonly used terms such as "disinformation" and "fake news". Next, we examine the psychology of correcting misinformation: what happens when someone is exposed to a fact-check, and what are the benefits and drawbacks of correcting misinformation once the damage is already done? Finally, we discuss how to build psychological resilience against misinformation through psychological "vaccines" or "inoculation": we look at the theoretical background of inoculation theory (dating back to the 1960s), its modern application within the context of online misinformation (in online games and educational videos), and future prospects and research avenues in the field.

1. DEFINING THE PROBLEM

A variety of terms are used interchangeably when discussing the spread of harmful, false and/or misleading information online, including but not limited to "misinformation", "disinformation", "malinformation" and "fake news" (13–16). In addition, scholars, policymakers, journalists and others employ various definitions for each of these terms. For example, a common definition of "fake news" is "fabricated information that mimics news media content in form, but not in organisational process or intent" (17). According to this definition, the feature that sets "fake news" apart from "real news" is factual veracity: fake news is fake only when it is "fabricated". However, some scholars disagree with this definition due to its relatively narrow scope: some fabricated news is mostly harmless (such as satirical news articles by The Onion or The Babylon Bee), whereas true information can be stripped of relevant context and presented in a misleading manner (van der Linden & Roozenbeek, 2020). To give an example (van der Linden & Roozenbeek, in press): while objectively false stories about COVID-19, such as the non-existent link between 5G radiation and symptoms of COVID-19 (6) are quite common, stories that are not quite false but highly misleading arguably have even more potential to do damage. For instance, the US Center for Disease Control (CDC) writes on its website that there is a "plausible causal relationship between the J&J/Janssen COVID-19 Vaccine and a rare and serious adverse event—blood clots with low platelets—which has caused deaths" (19). This may sound like the vaccine poses a serious risk, but the CDC also provides relative context for the likelihood of such an adverse event: "As of July 12, 2021, more than 12.8 million doses of the J&J/Janssen COVID-19 Vaccine have been given in the United States. CDC and FDA [Food and Drug Administration] identified 38 confirmed reports of people who got the J&J/Janssen COVID-19 Vaccine and later developed [thrombocytopenia syndrome]." As of mid-July 2021, there were thus 38 confirmed cases of vaccine-related thrombosis out of 12.8 million vaccinations (not all of whom died), which comes down to a probability of around 0.000003% or 1 in 337,000.

To give even more context, the odds of dying in a car (or other vehicle) accident in the United States in any given year are approximately 1 in 8,000 (20). In other words, while the risk of adverse events following COVID-19 vaccination does exist, it is relatively very small. Nonetheless, in April 2021, the *Chicago Tribune* (along with numerous other outlets) published a story with the following headline: *"A 'healthy' doctor died two weeks after getting a COVID-19 vaccine; CDC is investigating why"* (21). The cause of death was unknown at the time of publication (the *Tribune* later added an update, stating that there was not enough evidence to "rule out or confirm the vaccine was a contributing factor"), but without the relevant context, the story went viral and was shared millions of times (22).

The challenge with this and similar news items is that none of the information in the headline is factually incorrect; rather, it is the implicit message (namely that a healthy person, a medical professional no less, died from the COVID-19 vaccine) that makes the headline misleading.

Consequently, other definitions do not focus on veracity (i.e., true or false) but rather on *intent*: whether online content is harmful is determined by whether its producer is intentionally seeking to deceive or manipulate their audience (11). When paired with an organised influence campaign, such intentional manipulation is often called "disinformation" (23). However, it can be difficult to discern a person's intent (especially when considering the internet's anonymity); are they truly trying to deceive people, or are they merely advocating for a sincerely held belief? In addition, even intentionally manipulative messages spread by dishonest actors can hold some degree of truth value, and seek to fuel polarisation rather than spread false information (24, 25).

Defining "misinformation" is thus highly complex. For the purpose of this report, which primarily covers inoculation theory as a tool for mitigating misinformation, we will focus on whether online content is *manipulative*, rather than true or false. in other words, this report will evaluate whether content is epistemologically dubious because it uses one or more known manipulation techniques: emotionally manipulative language (26, 27), polarising language (28), conspiratorial reasoning (29), trolling (30), or logical fallacies (31, 32). The advantage of this approach is that it avoids the problems described above, with both factual veracity and psychological intent. In addition, these manipulation techniques can be identified by examining the wording or language use of a piece of online content, which opens up the possibility for technique- or logic-based inoculations (see section 4).

2. CORRECTING MISINFORMATION AND THE "CONTINUED INFLUENCE EFFECT"

Fact-checking initiatives have become an increasingly popular method for mitigating the spread of misinformation. This includes fact-checking initiatives, such as FullFact and Snopes, but also "on-demand" fact-checkers like Repustar, which have developed innovative tools for detecting and correcting misinformation in real time. There is extensive literature available on debunking misinformation, as scholars have sought to gain insight into the factors, contexts, and circumstances under which corrections are effective. A large group of over 20 experts recently published a step-by-step guide detailing the best debunking practices, called the "Debunking Handbook" (33), which was based on an earlier version from 2011.

However, despite these advances, there are several limitations to debunking that ensure that post-hoc corrections are unlikely to be sufficient on their own. **First, a study by Vosoughi and colleagues (34) found that false rumours can spread further, faster and deeper through social networks than information that was later fact-checked and rated true** (34), although later research found that this may not be the case under all circumstances (35). Consequently, false information may reach more people than verified information, especially in homogeneous environments or "echo chambers". In other words, fact-checks may be unlikely to reach the same people as the original misinformation.

Second, people who have been exposed to misinformation may continue to rely on it, even if it has been debunked – a phenomenon known as the "continued influence effect" (36, 37). As such, we cannot expect fact-checks to reliably and comprehensively undo the damage done by misinformation exposure; myths may linger in our memory networks even after they were shown to be false.

Third, repeated exposure to misinformation increases its fluency and familiarity and therefore has the occasional side effect of increasing people's belief in it (even while knowing it to be false) – a phenomenon known as the "illusory truth effect" (38). In other words, if a misleading story goes viral on social media and is posted by different sources, people may see it multiple times when scrolling through their feed. This, in turn, may make the story seem more reliable to the people who see it. Intentional influence (or disinformation) campaigns may be especially well-suited to exploit this phenomenon: coordinating the spread of a particular story through multiple

> *Fact-checks may be unlikely to reach the same people as the original misinformation; myths may linger in our memory networks even after they were shown to be false.*

sources may amplify not only its potential reach, but also how many times people will see the story being shared, thus increasing its potential perceived credibility.

Fourth, there is some evidence that people do not like being fact-checked and may respond negatively to debunking attempts. For example, when researchers replied to a sample of 2,000 Twitter users who had previously shared false political news with a link to fact-checking websites, they found that contacted users subsequently retweeted lower quality news with a higher partisan slant and language toxicity (39). These results point to the potential influence of politically motivated reasoning, which can in some cases override people's desire to be accurate (40).

For these reasons, debunking misinformation, while effective under some circumstances, is not enough on its own. It is therefore necessary to also consider methods of preventing misinformation from taking root in the first place. Within the field of psychology, this means focusing on building psychological resistance against manipulation attempts preemptively, with a view toward rendering them less effective.

3. INOCULATION THEORY AND PREBUNKING

In response to the shortcomings of the post-hoc correction methods described above, scholars have explored ways to pre-emptively debunk (or "prebunk") misinformation. The idea of developing a psychological "vaccine" against misinformation derives from a framework from the 1960s called inoculation theory (41-43). During the Vietnam War, the U.S. government became concerned about the prospect of its troops becoming brainwashed (or persuaded) by foreign propaganda. This concern prompted the social psychologist William McGuire to explore the idea of a "vaccine for brainwash". Drawing on the analogy of medical inoculations, McGuire proposed that rather than bombarding people with more supportive facts, pre-emptively exposing people to a weakened dose of a specific persuasive [manipulative] argument could confer psychological resistance against future exposure to persuasive attacks, much like a medical vaccine confers physiological resistance against future infection (44). Over the years, inoculation treatments came to feature two core components: 1) a forewarning of an impending attack on one's beliefs, and 2) a pre-emptive refutation of the persuasive argument, also called a "prebunk" (43, 45). Since then, a large volume of studies and meta-analyses has been conducted, establishing inoculation theory as a robust framework for countering unwanted persuasion (2, 46).

Although the original paradigm has proved highly replicable (46), for a long time it was never tested in the context that inspired McGuire's idea: brainwashing and propaganda. This began to change around 2017, when researchers started to apply inoculation theory within the modern context of online misinformation (32, 47). Van der Linden et al. (5), for example, looked at whether it is possible to "inoculate" people against misinformation about climate change. Study participants were shown an image that contained either a message about the scientific consensus regarding climate change ("97% of climate scientists have concluded that human-caused climate change is happening"), a misinformation message (a petition said to have been signed by more than 30,000 people claiming that there is "no convincing scientific evidence" that human-caused climate change is harmful), or both messages alongside each other.

In addition, some participants were shown either a general ("partial vaccine") or specific ("full vaccine") inoculation message before the misinformation. The general inoculation was worded as follows (the specific inoculation goes into more detail about why the petition containing misinformation was unreliable – for example, because it was signed by Charles Darwin, who is dead):

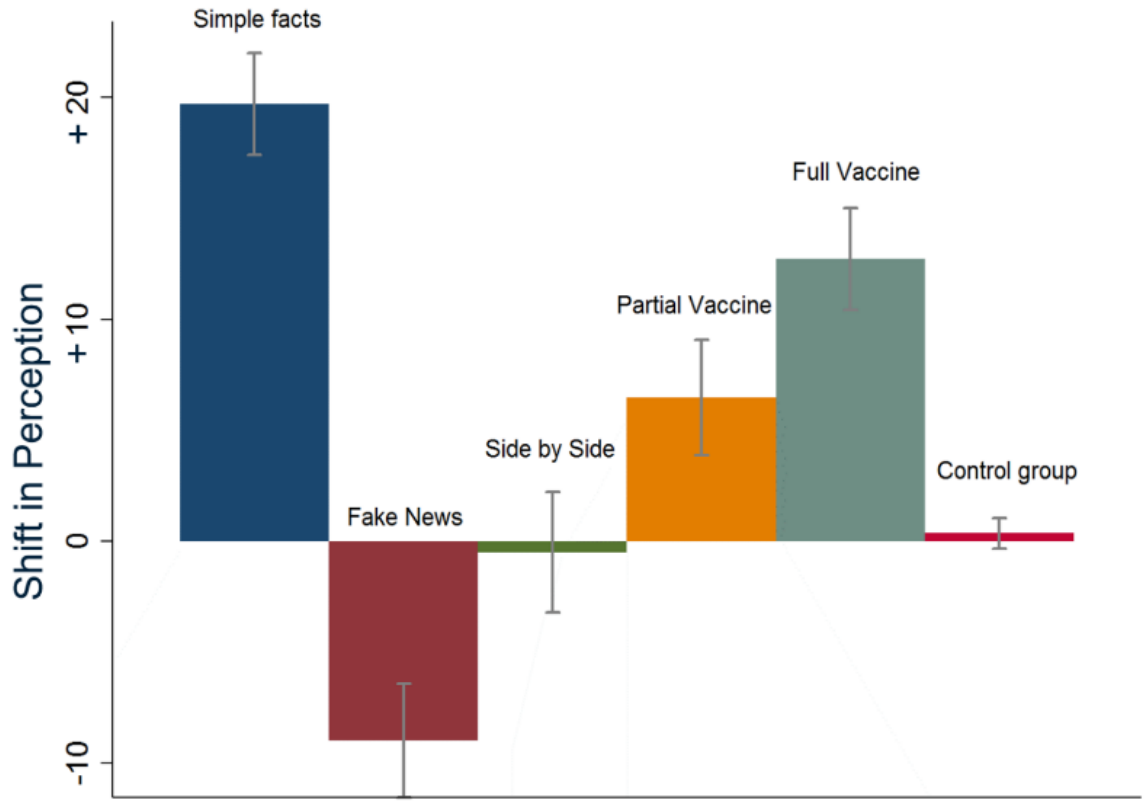

Figure 1. Shift towards or away from the scientific consensus about climate change after seeing a consensus message, "fake news", both (side by side), and a general ("partial vaccine") and specific ("full vaccine") inoculation message. Reprinted with permission from Van der Linden et al. (5).

> Nearly all climate scientists – 97% – have concluded that human-caused climate change is happening. Some politically-motivated groups use misleading tactics to try to convince the public that there is a lot of disagreement among scientists. However, scientific research has found that among climate scientists "there is virtually no disagreement that humans are causing climate change".

Van der Linden and colleagues then measured the extent to which participants' perception of the scientific consensus about climate change (namely, that it is happening and human-caused) shifted towards or away from the actual consensus after seeing the messages. The results are shown in Figure 1.

Figure 1 shows that both partial and full inoculation are effective at countering the effects of misinformation exposure; participants exposed to either inoculation message shifted their perception of the scientific consensus about climate change in the right direction (i.e., towards the 97% consensus), whereas participants who only saw the misinformation became more sceptical of the scientific consensus. Crucially, Van der Linden and colleagues also found both inoculation treatments to be effective across the political spectrum. This experiment thus showed the potential of inoculation theory as a tool for reducing susceptibility to online misinformation.

4. TECHNIQUE-BASED INOCULATION

While the aforementioned findings are promising, the inoculation treatments used in the study by Van der Linden et al. have one important limitation: they revolve around a single persuasive attack. This poses a limitation to the scalability of inoculation interventions, because it is not feasible to design and implement inoculations against every conceivable example of misinformation discoverable online. To address this limitation, researchers began exploring the feasibility of moving away from issue-based and towards logic- or technique-based inoculations (32, 48). As the name suggests, technique-based inoculations expose the manipulation techniques that are commonly used in online misinformation, such as floating conspiracy theories (i.e., blaming a small, secretive group of people with ill intentions for societal problems; see 49), the use of emotionally manipulative language to evoke strong emotions such as outrage or fear (26, 27), using language intended to fuel intergroup tensions and polarisation (28, 50), or artificially amplifying the reach of one's content through bots or fake "likes" (11). If people were to be inoculated against these techniques (rather than only against their usage in specific examples of misinformation), then they might become better able to recognise the use of such techniques in the content they see online. Such an approach circumvents the need to prebunk individual examples of misinformation, and since this approach tackles epistemologically dubious content (and therefore sidesteps the question of what counts as "fact"), may be less likely to be perceived as biased than a fact-check.

5. INOCULATION GAMES AND VIDEOS

Initial research into the feasibility of technique-based inoculation was promising: both Cook and colleagues (32) and Roozenbeek and Van der Linden (48) found that pre-emptively explaining and warning against common manipulation techniques subsequently reduced susceptibility to unseen misinformation. However, there was an open question with respect to the design of such interventions: how can you ensure that people voluntarily engage with the inoculation treatment? And, perhaps even more importantly, what is the longevity of the inoculation effect? After all, if people only benefit from a technique-based inoculation for a few seconds, as may be the case with other anti-misinformation interventions, such as "accuracy nudges", also known as accuracy primes (51), then the overall efficacy of the intervention may be limited.

Inoculation Games

As a first step, researchers designed a series of free online games aimed at reducing susceptibility to misinformation techniques. Games are a promising medium for inoculation interventions because of their potential entertainment value and volume of voluntary uptake, as well as the level of cognitive effort required to complete them. In effect, the more time and effort is spent learning about how misinformation techniques work, the more effective the intervention is likely to be. **The first game was Bad News (www.getbadnews.com), a choice-based browser game created by DROG and the University of Cambridge in which players take on the role of a fake news producer.** Over the course of 6 levels, each covering a single misinformation technique such as trolling, conspiratorial reasoning, or impersonating fake accounts, players grow from an anonymous social media user to a "fake news tycoon". A first evaluation of the game's efficacy with over 15,000 responses – assessed through a voluntary survey embedded within the game environment – showed that *Bad News* players find tweets containing a misinformation technique significantly less reliable after playing compared to before (52). Subsequent studies demonstrated the robustness of the inoculation effect conferred by the game: in a series of randomised controlled trials, researchers replicated the original finding that playing the game reduces the perceived reliability of misinformation, and also found that the game increases people's confidence in their ability to recognise misleading content (53). This is important because when people

are not confident in their own misinformation detection abilities, they can be easily persuaded. In addition, the inoculation effect was robust across different ages, education levels, and political ideologies, remained consistent in five different language versions of the *Bad News* game (54), and does not appear to suffer from serious item- or testing effects (55). In terms of the longevity of the effect, a longitudinal study showed that the reduction in the perceived reliability of misinformation after playing *Bad News* remained significant for a period of at least three months post-gameplay, provided participants were given regular reminders or "booster shots" of the misinformation techniques they learned about in the game (56).

Figure 2 shows a bar graph of the perceived reliability of misinformation over time for the group of participants that played *Bad News* and the control group, which played *Tetris*.

Since the launch of *Bad News*, several other inoculation games have been created, each of which covers a different domain of misinformation. *Harmony Square* (www.harmonysquare.game, developed by DROG, the U.S. Department of State's Global Engagement Center and the University of Cambridge) tackles disinformation and polarisation. The game's setting is Harmony Square, a peaceful community known for its pond swan and annual pineapple pizza festival.

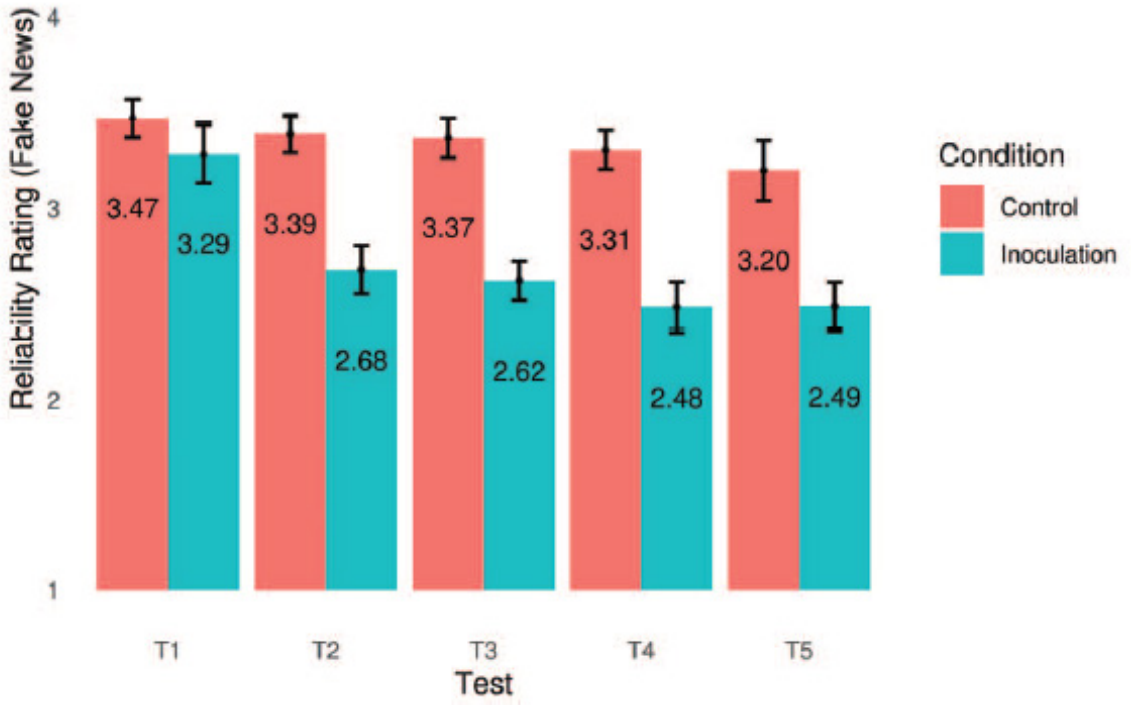

Figure 2. Perceived reliability of misinformation before the intervention (T1), immediately after the intervention (T2), and one (T3), five (T4) and thirteen (T5) weeks after, for the inoculation (Bad News) and control (Tetris) group. The figure shows that participants in the inoculation group continue to rate misinformation as significantly less reliable than the control group over time, with repeated reminders or "booster shots". Error bars show 95% confidence intervals. Reprinted with permission from Maertens et al. (56).

Players are hired as the "Chief Disinformation Officer" for a malign organisation and are tasked with mounting an influence campaign to drive the people of Harmony Square apart. Over the course of four levels, players reduce the square to metaphorical rubble while learning about how disinformation and trolling campaigns can be used to fuel intergroup polarisation (24, 28). The game is available in numerous languages, including Arabic, Bahasa, Czech, Dutch, English, French, German, and Russian. A large-scale randomised controlled study showed that playing *Harmony Square*, like *Bad News*, reduces the perceived reliability of misinformation and increases people's confidence in spotting it. In addition, players were significantly less likely to indicate willingness to share misinformation with people in their network than a control group (57).

Go Viral! (www.goviralgame.com, developed by DROG, the University of Cambridge and the UK Cabinet Office) focuses on misinformation related to COVID-19. This 5-minute game simulates the player's descent into an online echo chamber where misinformation is common. Across three levels, players learn about the use of emotionally manipulative language, the use of fake experts to lend credibility to misinformation, and the use of conspiratorial reasoning to sow doubt. In a study with more than 3,500 participants, researchers found that playing *Go Viral!* confers a similar inoculation effect to *Bad News* and *Harmony Square*, in that players were significantly better at discerning COVID-19 misinformation from non-misinformation, were more confident in their ability to do so, and were less likely to indicate willingness to share COVID-19 misinformation with others. These effects were similar across three different language versions of the game (English, French and German), and remained detectable for at least one week post-gameplay. In addition, the researchers compared *Go Viral!* to a set of COVID-19 misinformation infographics created by UNESCO, which were designed to serve as a prebunking tool against misinformation and have the advantage of being easily implementable in social media environments. They found that both the game and the infographics were effective at conferring psychological resistance against COVID-19 misinformation. However, the game yielded descriptively larger effect sizes, and the inoculation effect remained detectable for a longer period of time (58).

Cranky Uncle (www.crankyuncle.com, developed by John Cook at Monash University in collaboration with creative agency Autonomy) is a free app-based game that focuses on climate misinformation. It covers 14 techniques of science denial ranging from fake experts to cherry-picking and a variety of different logical fallacies (such as red herrings, false equivalence, and the strawman fallacy). The game features gameplay elements, such as interactive quizzes and offers feedback to engage players, rewarding longer gameplay while building stronger resilience against climate misinformation. *Cranky Uncle* was designed for classroom use, and the website features a teacher's guide, which explains the application of the game as an educational tool. Preliminary data collected through an (unpublished) in-game critical thinking quiz, in which participants had to identify what type

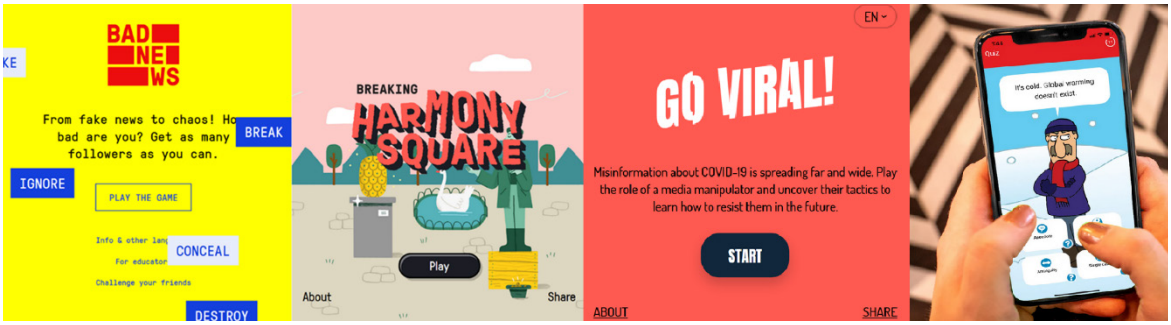

Figure 3. Bad News (www.getbadnews.com), Harmony Square (www.harmonysquare.game), Go Viral! (www.goviralgame.com) and Cranky Uncle (www.crankyuncle.com) game environments.

of reasoning fallacy (if any) is used in a series of stimuli both before and after playing *Cranky Uncle*, showed a significant improvement in critical thinking performance post-gameplay. Figure 3 shows the *Bad News, Harmony Square*, *Go Viral!* and *Cranky Uncle* game environments.

INOCULATION VIDEOS

In mid-2020, researchers from the University of Bristol and the University of Cambridge, in collaboration with Google Jigsaw, began exploring whether technique-based inoculation could be achieved using short videos instead of games. The advantages that videos have over games is that they are less time consuming, require less commitment from participants, and can be more easily implemented as advertisements on video streaming platforms (such as YouTube) and social media, or as part of educational classes and workshops. In other words, technique-based inoculation videos offer potential for a significant increase in the scalability of inoculation interventions.

The researchers developed five inoculation videos, each about 1.5 minutes in length, covering five manipulation techniques commonly found in online misinformation: the use of emotionally manipulative language (26), incoherence (59), false dichotomies or false dilemmas (60), scapegoating (61), and ad hominem attacks (62). In a large study with more than 5,000 participants, the researchers then tested whether watching an inoculation video improved people's ability to recognise the use of a manipulation technique in social media content; increased their confidence in their ability to do so; reduced the perceived trustworthiness of manipulative social media content; and improved the quality of people's decisions of whether to share content with their network, compared to a control group that watched an unrelated video of similar length. They found strong support for all of these hypotheses. Figure 4 shows the results for "technique discernment", i.e., the ability to detect whether a manipulation technique is used in a social media post (31). Crucially, researchers also found that the inoculation effect was robust across the

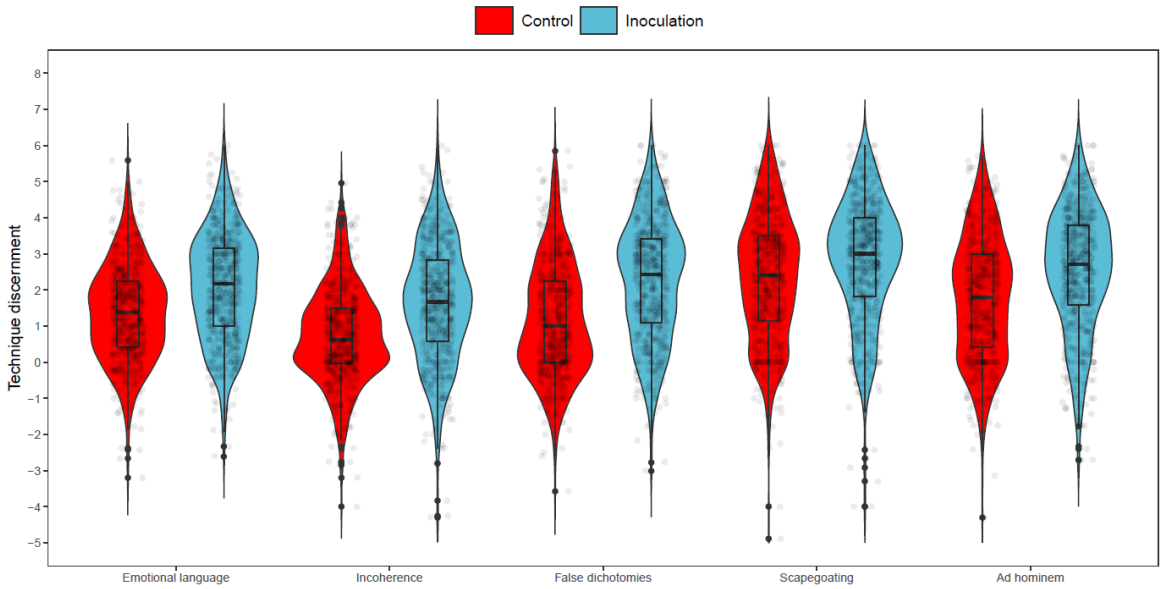

Figure 4. Technique discernment (i.e., the ability to discern manipulative from non-manipulative social media content) for the inoculation (treatment) group and control group, for each of the five inoculation videos. The figure shows that technique discernment is significantly higher in the inoculation group (blue) than the control group (red) for all five videos. Reprinted with permission from Roozenbeek et al. (31).

political spectrum, which is important because political partisanship can be a moderator of intervention efficacy, as appears to be the case with the aforementioned accuracy primes (63).

In a subsequent study (yet unpublished), the researchers tested whether the videos could be shortened to +-30 seconds, to broaden their accessibility as advertisements on social media and video sharing platforms. They found that the shortened videos were approximately equally effective as the longer ones. The videos, along with additional information about the previously described studies, can be found at www.inoculation.science.

6. FUTURE DIRECTIONS

The research described prior demonstrates that inoculation theory is a useful framework within the context of countering online misinformation. Inoculation interventions, including games, videos and infographics, have been shown to be effective at conferring psychological resistance against future unwanted persuasion attempts. In addition, inoculation interventions boost attitudinal certainty (i.e., confidence) about detecting misinformation on social media, and, crucially, reduces their self-reported willingness to share misinformation with others in their network. Furthermore, these effects are (largely) not moderated by covariates, such as political partisanship, age, and education, indicating that the interventions are effective across broad population groups. Nonetheless, several avenues for future research remain to be explored (64).

First, although some research has been done into the longevity of the inoculation effect (56, 58, 65), it remains unknown what exactly this "decay curve" looks like; how quickly do the effects dissipate over time (e.g., one week, two weeks, one month or even longer), and when exactly so-called "booster shots" should be administered to retain maximum efficacy is an important topic of ongoing research. At the time of writing, this question is being investigated for the above-mentioned inoculation videos.

Second, inoculation interventions perform very well in randomised controlled trials and within-intervention surveys. However, the extent to which such interventions influence people's engagement with misinformation in the real world remains unknown. For example, does playing an inoculation game or watching a video influence the quality of the content that people share on social media? Field studies are needed to bring more insight into how lab or simulated performance translates into real-world efficacy.

Third, how inoculation interventions perform compared to other anti-misinformation interventions such as accuracy primes (51, 63), introducing friction in people's social media environments (66), and media literacy interventions (67, 68) is currently unknown, as a direct comparison is yet to be done. Furthermore, whether deploying various interventions simultaneously (or in a complementary manner) yields a compounding effect in terms of reducing the spread of misinformation is a question awaiting further exploration.

Finally, the ultimate goal of psychological inoculation is herd immunity: what percentage of an online community needs to be "vaccinated", at what rate, and for how long, in order for sufficient immunity to be conferred?

And how do we respond to new misinformation methods and narratives? Computational models using the experimental effects obtained from the interventions described above are currently being designed to simulate population-level estimates for achieving psychological herd immunity against misinformation. After all, if enough people are vaccinated and have developed psychological antibodies, misinformation is less likely to be spread.

REFERENCES

1. S. Lewandowsky, U. K. H. Ecker, J. Cook, Beyond misinformation: Understanding and coping with the "Post-Truth" era. *J. Appl. Res. Mem. Cogn.* **6**, 353–369 (2017).
2. S. Lewandowsky, S. van der Linden, Countering Misinformation and Fake News Through Inoculation and Prebunking. *Eur. Rev. Soc. Psychol.*, 1–38 (2021).
3. N. F. Johnson, N. Velásquez, N. J. Restrepo, R. Leahy, N. Gabriel, S. El Oud, M. Zheng, P. Manrique, S. Wuchty, Y. Lupu, The online competition between pro- and anti-vaccination views. *Nature* (2020), doi:10.1038/s41586-020-2281-1.
4. B. L. Hoffman, E. M. Felter, K.-H. Chu, A. Shensa, C. Hermann, T. Wolynn, D. Williams, B. A. Primack, It's not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook. *Vaccine*. **37**, 2216–2223 (2019).
5. S. van der Linden, A. Leiserowitz, S. Rosenthal, E. Maibach, Inoculating the Public against Misinformation about Climate Change. *Glob. Challenges*. **1**, 1600008 (2017).
6. D. Jolley, J. L. Paterson, Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *Br. J. Soc. Psychol.* (2020).
7. BBC News, Ofcom: Covid-19 5G theories are "most common" misinformation. *www.bbc.co.uk* (2020), (available at https://www.bbc.co.uk/news/technology-52370616).
8. F. Vasudeva, N. Barkdull, WhatsApp in India? A case study of social media related lynchings. *Soc. Identities*. **26**, 574–589 (2020).
9. A. Warofka, An Independent Assessment of the Human Rights Impact of Facebook in Myanmar. *Facebook* (2018), (available at https://about.fb.com/news/2018/11/myanmar-hria/).
10. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election. *Science (80-.).* **363**, 374–378 (2019).
11. S. van der Linden, J. Roozenbeek, in *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*, R. Greifeneder, M. Jaffé, E. Newman, N. Schwarz, Eds. (Psychology Press, London, 2020).
12. L. Rainie, J. Anderson, J. Albright, "The Future of Free Speech, Trolls, Anonymity,and Fake News Online" (2017), (available at https://www.elon.edu/u/imagining/wp-content/uploads/sites/964/2019/07/Pew-and-Elon-University-Trolls-Fake-News-Report-Future-of-Internet-3.29.17.pdf).
13. M. D. Molina, S. S. Sundar, T. Le, D. Lee, "Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. *Am. Behav. Sci.* **65**, 180–212 (2019).
14. E. K. Vraga, L. Bode, Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation. *Polit. Commun.* **37**, 136–144 (2020).
15. S. Iyengar, D. S. Massey, Scientific communication in a post-truth society. *Proc. Natl. Acad. Sci.* **116**, 7656–7661 (2018).

16. T. Zerback, F. Töpfl, M. Knöpfle, The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media Soc.* **23**, 1080–1093 (2021).
17. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The science of fake news. *Science (80-.).* (2018), doi:10.1126/science.aao2998.
18. S. van der Linden, J. Roozenbeek, in *The Social Science of the COVID-19 Pandemic: A Call to Action for Researchers*, M. K. Miller, Ed. (Oxford University Press, Oxford, 2022).
19. Center for Disease Control and Prevention, Selected Adverse Events Reported after COVID-19 Vaccination. *www.cdc.gov* (2021), (available at https://www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/adverse-events.html).
20. Insurance Information Institute, Facts + Statistics: Mortality risk. *www.iii.org* (2021), (available at https://www.iii.org/fact-statistic/facts-statistics-mortality-risk).
21. A. Borgya, A 'healthy' doctor died two weeks after getting a COVID-19 vaccine; CDC is investigating why. *Chicago Trib.* (2021), (available at https://www.chicagotribune.com/coronavirus/fl-ne-miami-doctor-vaccine-death-20210107-afzysvqqjbgwnetcy5v6ec62py-story.html).
22. M. Parks, Few Facts, Millions Of Clicks: Fearmongering Vaccine Stories Go Viral Online. *NPR* (2021), (available at https://www.npr.org/2021/03/25/980035707/lying-through-truth-misleading-facts-fuel-vaccine-misinformation?t=1621596471521&t=1622544913288).
23. D. Freelon, C. Wells, Disinformation as Political Communication. *Polit. Commun.* **37**, 145–156 (2020).
24. C. A. Bail, B. Guay, E. Maloney, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, A. Volfovsky, Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proc. Natl. Acad. Sci.* **117**, 243–250 (2020).
25. E. Ferrara, Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *CoRR*. **abs/1707.0** (2017) (available at http://arxiv.org/abs/1707.00086).
26. W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, J. J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl. Acad. Sci.* **114**, 7313–7318 (2017).
27. M. Berriche, S. Altay, Internet users engage more with phatic posts than with health misinformation on Facebook. *Palgrave Commun.* **6**, 1–9 (2020).
28. A. Simchon, W. J. Brady, J. J. Van Bavel, Troll and Divide: The Language of Online Polarization. *PsyArxiv Prepr.* (2021), doi:10.31234/osf.io/xjd64.
29. S. Lewandowsky, G. E. Gignac, K. Oberauer, The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLoS One*. **8**, 1–11 (2013).
30. M. D. Griffiths, Adolescent trolling in online environments: a brief overview. *Educ. Heal.* **32**, 85–87 (2014).
31. J. Roozenbeek, S. van der Linden, B. Goldberg, S. Lewandowsky, Scaling psychological inoculation against misinformation techniques. *Sci. Adv.* (2021).
32. J. Cook, S. Lewandowsky, U. K. H. Ecker, Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One*. **12**, 1–21 (2017).

33. S. Lewandowsky, J. Cook, U. K. H. Ecker, D. Albarracín, M. A. Amazeen, P. Kendeou, D. Lombardi, E. J. Newman, G. Pennycook, E. Porter, D. G. Rand, D. N. Rapp, J. Reifler, J. Roozenbeek, P. Schmid, C. M. Seifert, G. M. Sinatra, B. Swire-Thompson, S. van der Linden, E. K. Vraga, T. J. Wood, M. S. Zaragoza, "The Debunking Handbook 2020" (2020), , doi:10.17910/b7.1182.
34. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science (80-.).* **359**, 1146–1151 (2018).
35. M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The COVID-19 social media infodemic. *Sci. Rep.* **10**, 16598 (2020).
36. U. K. H. Ecker, S. Lewandowsky, M. Chadwick, Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cogn. Res. Princ. Implic.* **5**, 41 (2020).
37. S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, J. Cook, Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychol. Sci. Public Interes.* **13**, 106–131 (2012).
38. L. Fazio, N. M. Brashier, B. K. Payne, E. J. Marsh, Knowledge does not protect against illusory truth. *J. Exp. Psychol. Gen.* **144**, 993–1002 (2015).
39. M. Mosleh, C. Martel, D. Eckles, D. G. Rand, in *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.
40. J. J. Van Bavel, A. Pereira, The Partisan Brain: An Identity-Based Model of Political Belief. *Trends Cogn. Sci.* **22**, 213–224 (2018).
41. W. J. McGuire, D. Papageorgis, The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *J. Abnorm. Soc. Psychol.* **62**, 327–337 (1961).
42. W. J. McGuire, Inducing resistance against persuasion: Some Contemporary Approaches. *Adv. Exp. Soc. Psychol.* **1**, 191–229 (1964).
43. J. Compton, in *The SAGE Handbook of Persuasion: Developments in Theory and Practice*, J. P. Dillard, L. Shen, Eds. (SAGE Publications, Inc., Thousand Oaks, ed. 2, 2013; http://sk.sagepub.com/reference/hdbk_persuasion2ed), pp. 220–236.
44. W. J. McGuire, A vaccine for brainwash. *Psychol. Today*. **3**, 36–64 (1970).
45. J. Compton, M. Pfau, Inoculation Theory of Resistance to Influence at Maturity: Recent Progress In Theory Development and Application and Suggestions for Future Research. *Ann. Int. Commun. Assoc.* **29**, 97–145 (2005).
46. J. A. Banas, S. A. Rains, A Meta-Analysis of Research on Inoculation Theory. *Commun. Monogr.* **77**, 281–311 (2010).
47. S. van der Linden, E. Maibach, J. Cook, A. Leiserowitz, S. Lewandowsky, Inoculating against misinformation. *Science (80-.).* **358**, 1141–1142 (2017).
48. J. Roozenbeek, S. van der Linden, The fake news game: actively inoculating against the risk of misinformation. *J. Risk Res.* **22**, 570–580 (2018).
49. K. M. Douglas, R. M. Sutton, Why conspiracy theories matter: A social psychological analysis. *Eur. Rev. Soc. Psychol.* **29** (2018), doi:10.1080/10463283.2018.1537428.
50. S. Rathje, J. J. Van Bavel, S. van der Linden, Outgroup animosity drives engagement on social media. *Proc. Natl. Acad. Sci.* (2021).

51. J. Roozenbeek, A. L. J. Freeman, S. van der Linden, How accurate are accuracy nudges? A pre-registered direct replication of Pennycook et al. (2020). *Psychol. Sci.* **32**, 1–10 (2021).
52. J. Roozenbeek, S. van der Linden, Fake news game confers psychological resistance against online misinformation. *Humanit. Soc. Sci. Commun.* **5**, 1–10 (2019).
53. M. Basol, J. Roozenbeek, S. van der Linden, Good news about Bad News: Gamified inoculation boosts confidence and cognitive immunity against fake news. *J. Cogn.* **3(1)**, 1–9 (2020).
54. J. Roozenbeek, S. van der Linden, T. Nygren, Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy Sch. Misinformation Rev.* **1** (2020), doi:10.37016//mr-2020-008.
55. J. Roozenbeek, R. Maertens, W. McClanahan, S. van der Linden, Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation. *Educ. Psychol. Meas.* **81**, 340–362 (2021).
56. R. Maertens, J. Roozenbeek, M. Basol, S. van der Linden, Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *J. Exp. Psychol. Appl.* **27**, 1–16 (2021).
57. J. Roozenbeek, S. van der Linden, Breaking Harmony Square: A game that "inoculates" against political misinformation. *Harvard Kennedy Sch. Misinformation Rev.* **1** (2020), doi:10.37016/mr-2020-47.
58. M. Basol, J. Roozenbeek, M. Berriche, F. Uenal, W. McClanahan, S. van der Linden, Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data Soc.* **8** (2021), doi:10.1177/20539517211013868.
59. S. Lewandowsky, J. Cook, E. A. Lloyd, The "Alice in Wonderland" mechanics of the rejection of (climate) science: simulating coherence by conspiracism. *Synthese*. **195**, 175–196 (2016).
60. K. Escandón, A. L. Rasmussen, I. Bogoch, E. j Murray, K. Escandón, J. Kindrachuk, COVID-19 and false dichotomies — A nuanced review of the evidence regarding public health, COVID-19 symptomatology, SARS-CoV-2 transmission, masks, and reinfection. *OSF Prepr.* (2020), doi:10.31219/osf.io/k2d84.
61. H. Hansen, in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. (Metaphysics Research Lab, Stanford University, Fall 2017., 2017).
62. D. Walton, *Ad Hominem Arguments* (The University of Alabama Press, Tuscaloosa and London, 1998).
63. S. Rathje, J. Roozenbeek, C. Steenbuch Traberg, J. J. Van Bavel, S. van der Linden, Partisan differences in the effectiveness of priming accuracy. *Nat. Matters Aris.* (2021).
64. S. van der Linden, J. Roozenbeek, R. Maertens, M. Basol, O. Kácha, S. Rathje, C. Steenbuch Traberg, How can psychological science help counter the spread of fake news? *Span. J. Psychol.* **24**, 1–9 (2021).
65. R. Maertens, F. Anseel, S. van der Linden, Combatting climate change misinformation: longevity of inoculation and consensus messaging effects. *J. Environ. Psychol.* **70** (2020), doi:10.1016/j.jenvp.2020.101455.
66. L. Fazio, Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Misinformation Rev.* **1** (2020), doi:10.37016/mr-2020-009.
67. T. Nygren, M. Guath, Swedish teenagers' difficulties and abilities to determine digital news credibility. *Nord. Rev.* **40**, 23–42 (2019).
68. A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, N. Sircar, *Proc. Natl. Acad. Sci.*, in press, doi:10.1073/pnas.1920498117.

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org