Accepted for publication in *European Psychologist* (20 Sept 2022)

Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions

Jon Roozenbeek[1], Eileen Culloty[2], & Jane Suiter[2]

[1] Department of Psychology, School of the Biological Sciences, University of Cambridge, Cambridge, United Kingdom.

[2] School of Communications, Dublin City University, Dublin, Ireland.

Abstract: Developing effective interventions to counter misinformation is an urgent goal, but it also presents conceptual, empirical, and practical difficulties, compounded by the fact that misinformation research is in its infancy. This paper provides researchers and policymakers with an overview of which individual-level interventions are likely to have an influence on the spread of, susceptibility to, or impact of misinformation. We review the evidence for the effectiveness of four categories of interventions: boosting (psychological inoculation, critical thinking, and media and information literacy); nudging (accuracy primes and social norms nudges); debunking (fact-checking); and automated content labelling. In each area, we assess the empirical evidence, key gaps in knowledge, and practical considerations. We conclude with a series of recommendations for policymakers and tech companies to ensure a comprehensive approach to tackling misinformation.

Contents

Introduction

Misinformation is a significant societal problem that has become an increasingly popular topic among researchers, policymakers, journalists, and the wider public. We define misinformation as any kind of false or misleading information. The latter does not necessarily have to be factually incorrect, but instead may distort facts, be stripped of relevant context, or use a logical fallacy (Roozenbeek & van der Linden, 2022). In online environments, misinformation can appear in the form of news stories and social media content, and can be spread deliberately, accidentally, or without malicious intent[1].

The spread of misinformation is implicated in the resurgence of vaccine-preventable diseases, the subversion of political norms, and the amplification of social divisions (Au et al., 2021; Azzimonti & Fernandes, 2018; Loomba et al., 2021). Within each issue domain, there exist striking gaps in public understanding of these issues and concerted efforts to manipulate public opinion (Lewandowsky et al., 2017). The prevalence of misinformation, particularly online, is therefore increasingly viewed as a crisis that demands urgent action (Farkas & Schou, 2020).

However, research into how misinformation spreads from person to person (Cinelli et al., 2020; Del Vicario et al., 2016; Zollo et al., 2015), the determinants of misinformation susceptibility (Ecker, Lewandowsky et al., 2022; Pennycook & Rand, 2019; Roozenbeek, Maertens et al., 2022; Van Bavel et al., 2021), and the design and testing of interventions to counter misinformation (Cook et al., 2017; Guess et al., 2020; Lee, 2018) is growing rapidly, but remains in its infancy. A Web of Science (https://webofscience.com/) search for "misinformation" shows that the topic exploded in popularity among researchers only after the 2016 US presidential elections, rising from 43 academic publications per year in 2000, to 231 in 2015, and to 1,925 in 2021.

As policymakers, tech companies, news providers, educators, and other actors are tasked with developing responses to tackle the problem, it is of key importance to assess the evidence base for existing types of anti-misinformation interventions. However, there are knowledge gaps regarding their influence, impact, and effectiveness. These knowledge gaps are compounded by the fact that most research is concentrated in Europe and North America, as well as by the broader challenges of researching online misinformation (Badrinathan, 2021; van der Linden, 2022). For example, there are conceptual challenges surrounding the definition of the problem (Freelon & Wells, 2020; Kapantai et al., 2021; Traberg, 2022), practical challenges arising from the scale of online content distribution (Traberg et al., 2022), and ethical challenges relating to interventions in free, legal speech (Nuñez, 2020).

Interventions can take place at the system level or at the individual level (Chater & Loewenstein, 2022; Kozyreva et al., 2020). System-level responses to tackling

[1] We acknowledge the diversity in the various definitions of "misinformation", "malinformation", "disinformation", "fake news", "false news", "propaganda", and other similar terms (Krause et al., 2020; Lazer et al., 2018; Roozenbeek & van der Linden, 2022; Tandoc et al., 2018). Our definition of misinformation is deliberately broad to be inclusive of these various terms, and to incorporate not only false but also misleading/manipulative content (Altay et al., 2021). See Freelon and Wells (2020) and Kapantai et al. (2021) for further discussion.

misinformation range from data-sharing proposals between tech companies and researchers, to laws prohibiting the spread of misinformation (Nuñez, 2020) and regulatory proposals for social media platforms' algorithms (Khan, 2021; Ulbricht & Yeung, 2022). These are perhaps the most controversial form of intervention given the potentially adverse consequences for freedom of expression and media freedom (Bontcheva et al., 2020). Problematically, authoritarian and proto-authoritarian states have implemented laws prohibiting the spread of "false" information, which are often used to target individuals who are critical of the authorities (International Press Institute, 2022). For an overview of the actions governments around the world have taken to tackle misinformation, see Funke and Flamini (2018).

This paper serves to provide researchers, tech companies, and policymakers with an understanding of the extent to which *individual-level* interventions are likely to influence the spread of, susceptibility to, or impact of misinformation. Kozyreva et al. (2020) define four entry points for policy interventions aimed at tackling digital challenges: *laws and ethics* (such as regulations and ethical guidelines); *technology* (such as automated harmful content detection); *education* (e.g., media and information literacy); and *psychology/behavioural sciences* (boosting, nudging, and technocognition). Here, we employ a modified version of Kozyreva et al.'s (2020) categorisation of educational, psychological, and behavioural interventions. We discuss four intervention categories: *Boosting* interventions (psychological inoculation, critical thinking, and media/digital literacy trainings), which seek to improve relevant competences and increase cognitive resistance; *nudging* interventions (accuracy prompts and social-norms interventions), which aim to guide people's behaviour through the design of choice architectures; *debunking* (including fact-checking); and *automated content labelling*. For each category, we assess the empirical evidence, identify key gaps in knowledge, and, where relevant, highlight practical and ethical implications. See Figure 1 for a flowchart of different types of system-level and individual-level anti-misinformation interventions.

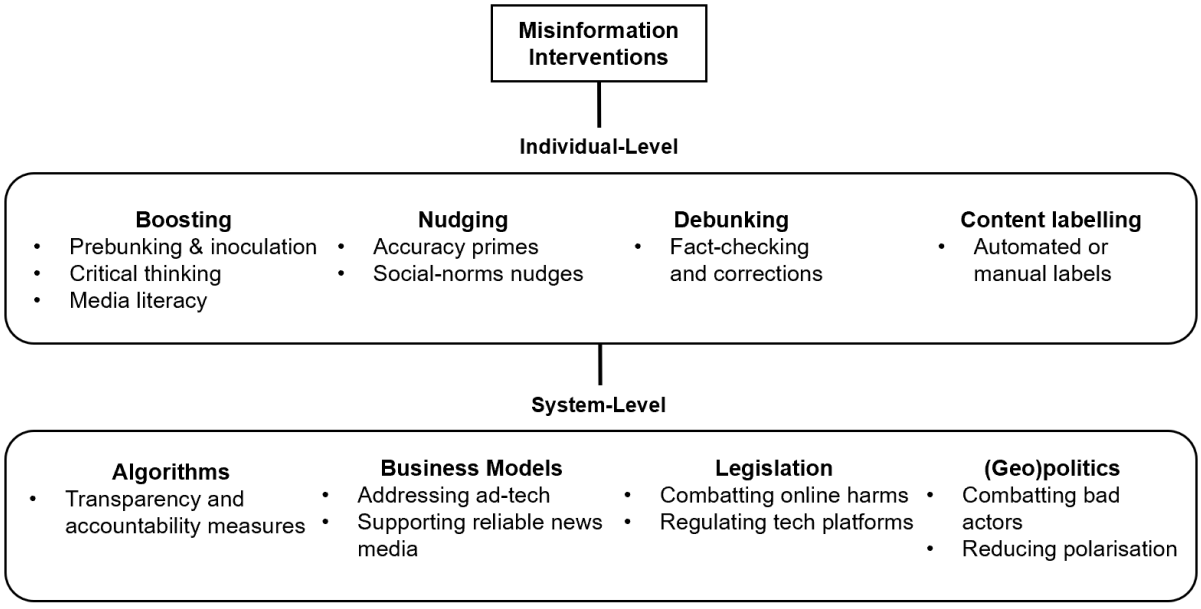

Figure 1. System-level and individual-level misinformation interventions.

Boosting Interventions

According to Hertwig and Grüne-Yanoff (2017, p. 974), the objective of boosts is "to improve people's competence to make their own choices", and that the focus of boosting is on "interventions that make it easier for people to exercise their own agency by fostering existing competences or instilling new ones", for example by improving people's ability to recognise microtargeted advertising (Lorenz-Spreen et al., 2021). Individuals are free to decide to disengage from boosts or not pay attention to them (as they are by nature non-mandatory and do not necessitate making changes to people's choice environment), making them unlikely to pose a significant ethical challenge or present major risks to the democratic process. For an overview, see https://scienceofboosting.org/.

Within the context of misinformation, boosting interventions tend to seek to reduce individual susceptibility to misinformation (van der Linden et al., 2021). Boosting interventions may not always be effective, for example because they fail to reach the relevant people (Zollo et al., 2017). In addition, although a large number of lab studies has been published in recent years investigating the effectiveness of various types of boosting interventions (see Kozyreva et al., 2020; Lorenz-Spreen et al., 2021), key open questions are to what extent effectiveness in the lab translates to the real world (Roozenbeek, van der Linden et al., 2022), and how boosting competences translates to changes in people's behaviour. We discuss three types of interventions that seek to improve people's ability to identify misinformation, and therefore can be said to fall under the boosting banner: prebunking (with a focus on psychological inoculation), critical thinking, and media and information literacy.

"Prebunking" and Psychological Inoculation

The challenges associated with correcting misinformation after it has spread (see the "Debunking" section below) have prompted researchers to explore how to prevent people from falling for and sharing misinformation in the first place (van der Linden et al., 2021). Such pre-emptive approaches to tackling misinformation are commonly referred to as pre-emptive debunking, or "prebunking" (Cook et al., 2017; Traberg et al., 2022).

Although several approaches to prebunking exist (Brashier et al., 2021; Cook et al., 2017; Lewandowsky & van der Linden, 2021; Tay et al., 2021), the most common framework for prebunking is inoculation theory (Compton, 2013; McGuire, 1961). People can build attitudinal resistance against future unwanted persuasion attempts by pre-emptively exposing them to a "weakened" dose of the unwanted persuasive argument (Compton et al., 2021). Inoculation treatments consist of two core components: 1) a warning of an impending attack on one's beliefs or attitudes (i.e., a forewarning of impending manipulation), and 2) a pre-emptive refutation of this upcoming manipulation attempt (Compton, 2013). A meta-analysis (Banas & Rains, 2010) found that inoculations are generally effective at increasing attitudinal resistance against unwanted persuasion, with a mean effect size of $d = 0.43$ (considered a moderate effect size).

There are two important distinctions within the context of inoculation interventions. The first is between *active* and *passive* inoculations (McGuire & Papageorgis, 1961; Traberg et al., 2022). With passive inoculation, people are provided with counterarguments against

the unwanted persuasion attempt, usually in the form of a short piece of text or a video. With active inoculation, people generate their own counterarguments, for example by playing a game.

The second distinction is between *issue-based* and *technique-based* inoculations. Issue-based inoculations seek to inoculate people against individual persuasive attacks or examples of misinformation, for instance about fair trade (Tay et al., 2021) or climate change (Maertens et al., 2020; van der Linden et al., 2017; Williams & Bond, 2020). In contrast, technique-based (also called logic-based) inoculations confer resistance against manipulation strategies or tactics such as logical fallacies, emotional manipulation, or conspiracy theories (Cook et al., 2017, 2018; Lewandowsky & Yesilada, 2021; Roozenbeek, van der Linden et al., 2022). Both approaches have their advantages: issue-based inoculations may be more effective than technique-based ones when it is known what misinformation people are likely to be exposed to in the near future (Zerback et al., 2021). Technique-based inoculations, on the other hand, have the benefit of applying to a wider range of misinformation, at the expense of specificity (Cook et al., 2017; 2018).

A range of inoculation interventions has been developed in recent years to counter misinformation (van der Linden, 2022). Passive inoculation interventions were found to successfully confer psychological resistance against misinformation about climate change (van der Linden et al., 2017; Williams & Bond, 2020) and COVID-19 (Basol et al., 2021), astroturfing comments (Zerback et al., 2021), vaccine conspiracies (Jolley & Douglas, 2017; Wong & Harrison, 2014), extremist propaganda (Braddock, 2019; Hughes et al., 2021), and "fake experts" (Cook et al., 2017)[2].

One line of research has explored the use of short, informative videos as inoculation interventions. Lewandowsky and Yesilada (2021) found that a short video inoculated individuals against both Islamic-extremist and Islamophobic content. Similarly, Hughes et al. (2021) and Piltch-Loeb et al. (2022)[3] developed and tested effective inoculation videos to counter extremist propaganda and vaccine misinformation, respectively. Roozenbeek, van der Linden et al. (2022) designed five videos, each inoculating people against a different manipulation technique. They found that watching such a video significantly reduced subsequent susceptibility to the use of these techniques in social media content, including in an ecologically valid field study on YouTube. The videos can be viewed on a website created by the researchers: https://www.inoculation.science/.

In terms of active inoculation interventions, recent research has focused primarily on inoculation games. Such games tend to inoculate people against a set of manipulation techniques commonly used in a particular domain where misinformation is common. Cook et al. (2022), for example, created *Cranky Uncle* (https://www.crankyuncle.com/), a free game (and phone app) that uses cartoons and humour to foster critical thinking and fight

[2] We note that the findings by Cook et al. (2017) were not replicated in a recent replication using a German sample (Schmid-Petri & Bürger, 2021), in the sense that, unlike in the original study, the inoculation did not have an effect on participants' climate-related attitudes, possibly due to lower baseline beliefs in climate misinformation.

[3] See this *YouTube* playlist for the videos developed by Piltch-Loeb et al. (2022): https://www.youtube.com/playlist?list=PLYPI-AWCOGj6oNUfi7ddbBEglL56h8I2C

misinformation about climate change. For further reading see Cook (2021) and Cook et al. (2022). Another example of an active inoculation intervention is *Bad News* (https://www.getbadnews.com), a browser game in which players strive to become a "fake news tycoon" by learning about six common manipulation techniques, such as trolling and ad hominem attacks. In a series of studies, *Bad News* was shown to significantly improve people's ability to identify misinformation techniques, and increase their confidence in their ability to do so (Basol et al., 2020; Roozenbeek, van der Linden, et al., 2020; Roozenbeek, Maertens, et al., 2021; Roozenbeek & van der Linden, 2019). Furthermore, these effects remained significant for several months post-gameplay if participants were given periodic reminders or "booster shots" (Maertens et al., 2021). Other inoculation games include *Harmony Square* (https://harmonysquare.game), about political disinformation and intergroup polarisation (Roozenbeek & van der Linden, 2020), *Go Viral!* (https://goviralgame.com), about COVID-19 misinformation (Basol et al., 2021), and *Radicalise*, about the manipulation strategies used by extremist organisations (Saleh et al., 2021).

Although inoculation interventions have several advantages, they also have numerous downsides. First, inoculations are generally somewhat lengthy, requiring both a forewarning and a pre-emptive refutation. The effectiveness of inoculations therefore relies on voluntary uptake (for example, not everyone wants to play a game). Second, it is not always possible to predict what misinformation people will be exposed to, and because inoculations generally require a degree of specificity to be effective (Zerback et al., 2021; Roozenbeek, Traberg, & van der Linden, 2022), they may not work very well if the discrepancy between the inoculation treatment and the misinformation is too large. Third, not much evidence is currently available on how inoculation interventions perform in the wild, for example on social media (but see the YouTube field study by Roozenbeek, van der Linden et al., 2022). Particularly when it comes to behaviour (e.g., people's news-sharing decisions), more research is needed to explore to what extent inoculation interventions are effective in real-world settings. Fourth, some inoculation interventions (notably "fake news games", see Modirrousta-Galian & Higham, 2022) appear to not only reduce belief in misinformation, but also reduce the perceived reliability of "real news". This appears to be a fairly common side effect of misinformation interventions in general (for example the media literacy tips by Guess et al., 2020, and the warning labels by Clayton et al., 2020). It is possible that some interventions induce a more general scepticism towards *all* information, although this phenomenon does not appear to apply to information that is obviously true (Roozenbeek & van der Linden, 2019; Basol et al., 2021; Modirrousta-Galian & Higham, 2022). Finally, if the "inoculator" is perceived as an untrustworthy actor, people may disregard the inoculation intervention. Like any intervention, inoculations risk becoming politicised (Traberg et al., 2022).

Critical Thinking

Critical thinking is typically defined as a higher-order skill that influences a person's ability to question assumptions, analyse arguments, and evaluate the quality of the information they encounter (Duron et al., 2006). There is disagreement about whether critical

thinking is a transferable skill that applies across domains or a domain-specific skill (Moore, 2014; Axelsson et al., 2021), and whether critical thinking is a skill (the *ability* to think critically) or a disposition (the *willingness* to think critically).

Various interventions to improve critical thinking have been tested. Lutzke et al. (2019), for instance, found a small effect for the effectiveness of reading a series of guidelines for evaluating online news, which improved individuals' ability to correctly evaluate the credibility of real and fake news about climate change on Facebook. A meta-analysis by Huber and Kuncel (2016) concluded that while university appears to foster a critical thinking disposition, specific interventions to improve critical thinking do not necessarily produce long-term incremental gains. However, many studies are small-scale or methodologically problematic; the evidence is therefore not strong enough to be conclusive (El Soufi & See, 2019; Todd & O'Brien, 2016). More large-scale, replicable, robust studies are required to advance the field.

From an ethical standpoint, boyd (2018) argues that critical thinking may be unhelpful if it encourages critical stance as a default. However, this view fails to distinguish between healthy scepticism and dysfunctional cynicism, whereby the latter is associated with lower trust in news media generally (Quiring et al., 2021). Moreover, the rhetoric of critical thinking may be adopted by those promoting conspiratorial beliefs who encourage people to "do your own research" and "ask questions" about, for example, the legitimacy of scientific evidence (Beene & Greer, 2021).

Media and Information Literacy

Since 2007, UNESCO has championed "media and information literacy"[4] as an umbrella concept that incorporates competences relating to media literacy, information literacy, news literacy, and digital literacy. Each of these literacies originally developed as a separate field, but the distinction between them is becoming increasingly blurred. With a specific focus on young people, *media literacy* has pursued the twin aims of protection (from the influence of media advertising, stereotypes, and bias) and empowerment (participating in media creation and self-expression; Hobbs, 2021). In contrast, *information literacy* puts emphasis on competences for finding and evaluating information and has traditionally been taught by librarians. Similar to media literacy, *news literacy* rests on the assumption that knowledge of news production practices will equip people to evaluate content more accurately (Tully et al., 2020). Finally, *digital literacy* involves the "necessary skills and competences to perform tasks and solve problems in digital environments" (Reddy et al., 2020). Within the context of misinformation, these fields often overlap.

Media and information literacy interventions are often conceived within formal education through the provision of teacher training and lessons on media and information literacy (Nygren & Guath, 2021). For example, researchers at the University of Uppsala have developed the News Evaluator Project (https://nyhetsvarderaren.se/in-english/), a series of free teaching materials aimed at boosting students' digital source criticism. Another example is Stanford University's Civic Online Reasoning initiative (https://cor.stanford.edu/), which

[4] See: https://www.unesco.org/en/communication-information/media-information-literacy

provides free, classroom-ready lessons and curricula about topics such as lateral reading (looking for information on other websites about a particular source), click restraint, and evidence evaluation. A growing body of research has found that such educational curricula are effective at increasing lateral reading and other strategies for navigating digital news environments (Craft et al., 2017; McGrew et al., 2019; Axelsson et al., 2021; Breakstone et al., 2021; Wineburg et al., 2022).

To improve the scalability of media and information literacy interventions outside of classroom settings, researchers have investigated how to deploy media and information literacy on social media (Tully et al., 2020). For example, Panizza et al. (2022) found that providing social media users with a pop-up that advised them on how to use lateral reading techniques subsequently increased use of such strategies. Another method for improving media literacy on social media involves literacy tips. A large-scale study evaluating the effectiveness of media literacy interventions in the United States and India found that providing people with media literacy tips to spot false and misleading content improved discernment of true and false news, although the intervention was only successful in India for a highly educated sample, but not for respondents from a largely rural area (Guess et al., 2020).

There are some notable challenges surrounding literacy-based interventions (some of which are shared with the critical thinking-focused interventions discussed above). First, such interventions have primarily focussed on children and young people through delivery via formal education institutions (Petranová et al., 2017). Less attention has been given to adults' media and information literacy needs and the kinds of interventions that might be effective for older age groups (Lee, 2018).

Second, although some studies from for example India (Badrinathan, 2021; Guess et al., 2020) and Pakistan (Ali & Qazi, 2021) are available, research testing media literacy interventions outside the Western world remains scarce. This is important because effective interventions in developed countries may not work as well elsewhere. Badrinathan (2021), for instance, found that a one-hour media literacy training conducted in India did not significantly improve participants' ability to identify misinformation.

Third, the design of media and information literacy interventions varies considerably. They utilise different literacy concepts and the intervention ranges from a one-off exposure to a module of lessons over many months. Moreover, researchers use a wide range of measures, which impedes comparisons between studies (Potter & Thai, 2016; Roozenbeek, Maertens et al., 2022). Such conceptual differences appear to matter. Based on a national sample of US citizens, a recent study suggests that information literacy - but not media, news, or digital literacies - significantly increases the likelihood of identifying misinformation stories (Jones-Jang et al., 2019).

Fourth, the impact of media and information literacy education is not always clear. There is a general lack of comprehensive evaluation data and "the longitudinal nature of both assessing and updating media literacy programs makes this a perennial struggle" (Bulger & Davison, 2018, p. 1). Regarding misinformation specifically, some studies find that exposure

to media and information literacy education predicts resilience to political misinformation (Kahne & Bowyer, 2017), but other studies caution that media literacy endows individuals with a false sense of confidence (Bulger & Davison, 2018). For their part, social media platforms already provide media literacy interventions for users[5], but they generally fail to provide any information about their uptake or impact (Culloty et al., 2021).

Nudging Interventions

Thaler and Sunstein (2008, p. 6) define nudges as "any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentive". Unlike boosts, which target competences, nudges thus target behaviour. Some scholars argue that people share misinformation on social media primarily because they fail to pay sufficient attention to accuracy, for example because social media environments can be distracting (Pennycook & Rand, 2019; 2021). Researchers have therefore proposed a range of nudging interventions, so-called "accuracy nudges" (or accuracy prompts), that intend to shift people's attention *towards* accuracy, thus reducing their propensity to share misinformation with others. Fazio (2020), for instance, found that asking people to pause for a few seconds to consider the accuracy of news headlines significantly reduced their willingness to share false news. Rathje et al. (2022a) found that motivating people to be as accurate as possible improved accuracy and reduced partisan bias in detecting false headlines.

Indeed, the potential advantages of the nudging approach are numerous: nudges are easy to implement on social media (for example, Twitter now asks people if they are sure they want to retweet an article if they have not yet read it), cost-efficient, and mostly non-intrusive. Furthermore, accuracy nudges do not require people to "opt-in" to the intervention, making them easily scalable.

Potential downsides of nudging include reactance against the intervention (for example, Mosleh et al., 2021, found that correcting people who had previously shared false news on Twitter decreased the quality and increased the toxicity of these users' subsequent retweets; an unintended backfire effect), and a reduced effect size when implemented in real-world environments compared to the lab (DellaVigna & Linos, 2022). From an ethical standpoint, Kozyreva et al. (2020) point out that low-cost nudges may displace support for high-cost (or high-effort) measures (this is arguably the case for boosting interventions as well, and individual-level interventions more generally; see Chater & Loewenstein, 2022). Within the context of misinformation, various types of nudges have gained prominence in recent years. In our discussion, we focus on the two most prominent ones: accuracy primes and social norms interventions.

[5] See for instance Facebook's Digital Literacy Library (https://www.facebook.com/safety/educators) and Twitter's ongoing media literacy collaboration with UNESCO (https://blog.twitter.com/en_sea/topics/events/2019/Media-and-information-literacy).

Accuracy Primes

The most well-known type of accuracy nudge is the accuracy prime, which consists of asking people to evaluate whether a single (usually non-political and non-partisan) headline is accurate[6]. Doing so subtly reminds them of the importance of sharing accurate content, which should then improve the quality of their subsequent news-sharing decisions. Several experimental studies have shown that accuracy primes improve subsequent sharing discernment. A recent meta-analysis of 14 lab-based accuracy prime studies (Pennycook & Rand, 2022) found that accuracy nudges significantly improved "sharing discernment" (a measure of the quality of sharing decisions; Epstein et al., 2021; Pennycook, McPhetres, et al., 2020; Pennycook et al., 2021), although the effect size is considered small within psychological research[7]. In terms of field studies, Pennycook et al. (2021) found that sending a group of Twitter users who had previously shared low-quality, conservative-leaning sources such as *Breitbart* a direct message, which asked them to evaluate the accuracy of a (non-political) headline, subsequently increased the quality of the content they shared on Twitter, primarily by prompting them to share more high-quality sources such as the *New York Times* and *CNN*.

However, several replications and re-analyses of accuracy prime studies have added nuance to these findings (Pennycook & Rand, 2022; Pretus et al., 2021; Rathje et al., 2022b; Roozenbeek, Freeman, et al., 2021). Pennycook and Rand (2022) report that accuracy primes had no effect on the quality of people's sharing decisions in 4 out of the 14 studies that were included in their meta-analysis. Roozenbeek, Freeman, et al. (2021) also initially failed to replicate the effect reported in an earlier study (Pennycook, McPhetres, et al., 2020), only finding a small effect after collecting a second round of data, at about 50% of the original study's effect size. They also noted initial evidence of rapid decay, in the sense that the impact of the accuracy prime occurred mostly shortly after exposure, and appeared to wear off quickly afterwards. Rapid effect decay is a known phenomenon in the priming literature (Branigan et al., 1999; Trammell & Valdes, 1992), but requires further investigation within the context of accuracy primes.

Accuracy primes also appear to be moderated by political partisanship. Pretus et al. (2021) found no effect of accuracy primes on two samples, one of US conservatives and one of Spanish far-right voters. This finding was further buttressed by Rathje et al. (2022b), who re-analysed six previously published accuracy prime studies and found that the priming effect was much smaller for (US) conservatives than liberals. This apparent moderation effect is especially important in light of the well-known finding that (US) conservatives and far-right supporters appear to share more misinformation (Garrett & Bond, 2021; Grinberg et al.,

[6] Although various names exist for this particular type of intervention (such as "accuracy nudge", "accuracy prompt" and "evaluation treatment"; see Pennycook & Rand, 2022), we here use "accuracy prime" to distinguish the single-headline evaluation treatment from other, similar nudging interventions that fall under the banner of accuracy prompts/nudges (see Epstein et al., 2021; Pennycook & Rand, 2022). In doing so, we follow Pennycook, McPhetres et al. (2020, p. 777), who note that this type of accuracy nudge "subtly primed [participants] to think about accuracy by being asked to rate the accuracy of a single […] news headline".

[7] Pennycook and Rand (2022) report a (most likely unstandardised) meta-analytic regression coefficient of b = 0.034, 95% CI [0.026, 0.043]. The meta-analytic Cohen's d was not reported.

2019). It is worth noting that Pennycook et al. (2021) did find an effect of accuracy primes on a predominantly conservative sample in their field study on Twitter[8].

Why this moderation effect exists (and there is some debate about this, see Pennycook & Rand, 2022) is less clear, but may have to do with the accuracy priming effect being smaller for more persuasive misinformation, and for people who generally rate misinformation as more accurate. In a 16-country study, Arechar et al. (2022) found that the accuracy prime's effect on the quality of people's sharing decisions in a given country was strongly correlated with the disconnect between the perceived accuracy of headlines and sharing intentions in that country; in other words, the priming effect was smaller (or non-significant) in countries where people generally rated misinformation as more accurate. In the United States, conservatives tend to rate misinformation as more accurate than liberals (see Roozenbeek, Maertens et al., 2022), which could signify a smaller disconnect between accuracy and sharing, thus reducing the priming effect.

Social-Norms Nudges

Rather than emphasising accuracy, social-norms nudges draw attention to partisan or societal norms around news-sharing behaviour in order to improve the quality of people's news-sharing decisions. Social-norms nudges are relatively understudied compared to accuracy primes and other types of nudging interventions. Some social-norms nudges focus on emphasising norms around sharing misinformation in a general sense. For example, Andı and Akesson (2021) found that warning people about the abundance of false information and telling them that "most responsible people think twice before sharing articles with their network" significantly reduced the proportion of people willing to share a false news article with others. Gimpel et al. (2021) report that exposing people to injunctive (what behaviour most people approve or disapprove of) but not descriptive (what other people do in certain situations) social norms increased the likelihood of participants reporting fake news posts on social media as misinformation; a combined approach (with both injunctive and descriptive norms) had the most substantial effect.

Other social-norms nudges seek to counter specific false or equivocal beliefs. Cookson et al. (2021) found that a social-norms intervention (giving people feedback about participants' belief in anti-vaccine conspiracy theories, how much they thought other people, in this case parents from the UK, believed such conspiracies, and UK parents' *actual* levels of conspiracy belief) significantly reduced personal belief in anti-vaccine conspiracies. Ecker, Sanderson et al. (2022) report that descriptive norms reduced belief in worldview-congruent equivocal claims (e.g., about the economic impact of refugees), although a descriptive norm plus a specific refutation proved to be the most effective.

Epstein et al. (2021), on the other hand, did not find an effect of emphasising either partisan or social norms on news-sharing intentions. However, they did find that a combined intervention (social norms + literacy tips or social norms + an importance prime, i.e., priming

[8] Pennycook et al. (2021) found that the largest pre-post difference in news sharing behaviour was a post-treatment *increase* in the sharing of mainstream/high-quality news sources such as the *New York Times* and *CNN*, more so than a post-treatment *decrease* in the sharing of misinformation/low-quality content. At a theoretical level, why accuracy primes might prompt conservative US Twitter users to share more *New York Times* headlines requires further elaboration.

people about the importance of sharing accurate news) positively affected the quality of sharing intentions. Overall, social-norms interventions have thus yielded promising, albeit preliminary results (and, importantly, are just as easy to implement on social media as accuracy primes), but are yet to be studied in real-world social media environments.

Debunking

Debunking misperceptions after they have spread is a popular approach to tackling misinformation. Initiatives such as Snopes, FullFact and StopFake abound, and some have large numbers of followers on social media. Debunking can be fact-based (i.e., correcting a specific misperception) or logic-based, focusing on the epistemic quality of the misinformation or the manipulation techniques used to mislead (Cook et al., 2017; Vraga et al., 2020; Vraga & Bode, 2020). Some technology companies, notably Meta (formerly Facebook), use debunking, drawing on both automated (Thorne & Vlachos, 2018; Guo et al., 2022) and human-centred methodologies to moderate content on their platforms. Debunking is almost synonymous with fact-checking, but not quite: one can fact-check a story and rate it as true, whereas debunking only pertains to misinformation.

Several debunking resources have been developed in recent years. For example, researchers with JITSUVAX (https://jitsuvax.info/) have created a free resource for countering both the specific arguments commonly used in vaccine misinformation, and the attitudinal roots that underlie these arguments. In 2020, a group of researchers published the *Debunking Handbook*, which summarises the current state of debunking research (Lewandowsky et al., 2020). The *Handbook* notes that debunking is most effective if the following procedure is followed:

1. Lead with the fact[9]; make it simple, concrete, and plausible, and ensure it fits with the story being debunked.
2. Warn the audience that they are about to see misinformation, and mention it only once.
3. Explain how the misinformation is misleading.
4. Finish by reinforcing the fact, and making sure that the fact provides a plausible alternative explanation to the misinformation.

Some years ago, researchers raised concerns about potential "backfiring effects" (corrections ironically strengthening people's belief in the original misinformation, see Nyhan & Reifler, 2010). The "illusory truth effect" states that misinformation is perceived as more accurate with repeated exposure (Ecker, Lewandowsky, et al., 2020; Fazio et al., 2015). This finding prompted concern that repeating the misinformation while correcting it may inadvertently reinforce people's belief in it (the so-called "familiarity backfire effect"). However, more recent research has shown that such backfire effects are not reliably observed, and the risk of debunking "side effects" therefore appears to be low (Swire-

[9] Swire-Thompson, Cook et al. (2021) argue that the specific correction format may play a limited role when correcting misinformation, and the *Debunking Handbook* also says that in cases where the facts are very nuanced, it may be better to lead with an explanation of why the myth is false rather than first explaining the fact.

Thompson et al., 2020, 2022; Wood & Porter, 2019). Overall, meta-analyses have concluded that corrective messages are generally effective at reducing belief in misinformation, although it is more difficult to correct misinformation about politics and marketing than about health (Chan et al., 2017; Walter et al., 2020; Walter & Murphy, 2018). In addition, debunking messages appear to be less effective when they are less detailed (Ecker, O'Reilly, et al., 2020; Paynter et al., 2019), and political beliefs and knowledge also appear to attenuate the effect (Walter & Murphy, 2018).

Nonetheless, there are several factors that limit the effectiveness of debunking. First, *who* is doing the debunking appears to matter a great deal. The perceived expertise and trustworthiness of the source affect how likely someone is to accept the correction (Benegal & Scruggs, 2018; Ecker & Antonio, 2021; Vraga & Bode, 2017; Guillory & Geraci, 2013). Furthermore, Margolin et al. (2017) found that Twitter users are significantly more likely to accept a correction by an account that they follow than a correction by a stranger, indicating that strong connections between fact-checkers and misinformation spreaders are key for the effectiveness of debunking.

Second, debunks do not appear to reach the same people as the original misinformation. Zollo et al. (2017), for example, found that debunking posts rarely penetrate conspiracy echo chambers on Facebook, and instead mostly reach users that prefer to consume science-focused content (who are unlikely to believe the misinformation to begin with). Hameleers and van der Meer (2019) showed that people engage more with fact-checks when they are congruent with prior attitudes, and avoid them when they are incongruent, indicating that people have a confirmation bias when deciding what fact-checks to engage with. Furthermore, Vosoughi et al. (2018) found that content that fact-checkers had rated as true reached far fewer people than content that they had rated as false.

Third, even if corrective messages reach the people who were exposed to the original misinformation, correcting the misinformation does not always completely undo people's belief in it, a phenomenon known as the "continued influence effect" (Ecker, Lewandowsky, et al., 2020; Walter & Tukachinsky, 2020). This is thought to occur because information that was previously encoded into memory can influence one's judgments, even when more recent information discredits it (Johnson & Seifert, 1994).

Fourth, fact-checkers are faced with several political challenges. Many of the most contentious political arguments are subjective and ethical in nature and, as such, are not reducible to objective facts (Coleman, 2018). Graves (2016) argues that fact-checkers themselves are at risk of becoming politicised. For example, deciding to fact-check a politician's claim can be perceived as picking a side, and places fact-checks at the centre of political debates. There is also a suggestion that people prone to conspiracy thinking may be resistant to debunking, which makes using mainstream sources of evidence to refute conspiratorial beliefs difficult (Hayes, 2006, p. 13). Thus, interventions oriented at delegitimising the sources that produce misinformation are key to reducing their impact (Ahmed et al., 2020). Another problem is some fact-checkers' dependency on large donors. Meta, for example, funds a large-scale third-party fact-checking programme, which has been criticised for lacking transparency and for having an outsize influence over what kind of content gets throttled on social media (BMJ, 2021; Nyhan, 2017). Thus, the effectiveness of fact-checking depends in part on the cooperation of large social media companies.

Finally, the literature is somewhat mixed when it comes to the relative effectiveness of prebunking and debunking. Brashier et al. (2021)[10] and Tay et al. (2021), for example, both report a descriptive difference in effect size in favour of debunking when comparing pre-emptive and post-hoc misinformation corrections (although it must be noted that this difference was not significant in both studies). Jolley and Douglas (2017), conversely, found that only prebunking was effective (and debunking was not) at countering the adverse effects of exposure to anti-vaccine conspiracy theories. Similarly, Grady et al. (2021) found that pre-warnings were descriptively more effective than debunking at discouraging belief in false political news headlines. However, a direct comparison between debunking and prebunking, although scientifically interesting, is subject to limitations. Because prebunking is preventative in nature (i.e., it occurs prior to exposure to misinformation), it is not usually possible to know exactly what misinformation to prebunk (as you cannot know what misinformation people will be exposed to in the future, or in what form). In other words, a prebunk and a debunk about the same misinformation will rarely (if ever) look exactly the same, and furthermore do not have the same goal (correcting a misperception versus preventing it from taking hold).

Automated Content Labelling

Online platforms have championed the use of automation to quickly label content at scale, increasingly rely on automated interventions to moderate the large volumes of content uploaded to their systems, and provide users with information to assess the credibility of information or sources (Alaphilippe et al., 2019). There are different types of content labels, for example fact-checks (e.g., "this article was rated false by independent fact-checkers"; see Brashier et al., 2021), general or specific content warnings (Clayton et al., 2020; Mena, 2019), and news credibility labels (Aslett et al., 2022). Automated content labelling typically relies on machine learning and neural network models to automate the content moderation process (Alaphilippe et al., 2019; Gorwa et al., 2020). Although specific techniques vary considerably, the overall aim is to classify content into problem categories (e.g., probable misinformation) or to match uploads against a database of problem content (e.g., known cases of misinformation); see Thorne and Vlachos (2018) and Guo et al. (2022) for an overview. Research into the effectiveness of content labels, such as smoking and alcohol warnings, has identified several necessary information processing steps (Conzola & Wogalter, 2001). The label must attract attention and maintain attention long enough so that all relevant information is noticed. The label must be understood, and the receiver's beliefs and attitudes may influence this comprehension. Finally, the label must motivate the receiver to adopt the suggested action. Overall, Bode and Vraga (2018) found that automated corrections and

[10] As noted by Lewandowsky and Yesilada (2021), the study by Brashier et al. (2021) deviated from conventional guidelines of de- and prebunking; in this study, participants were shown a label (saying "fact-checked and rated false/true") either before or after a false or true news headline. However, in contrast to other studies (Cook et al., 2017; Jolley & Douglas, 2017) and best practice guidelines (Lewandowsky et al., 2020), no refutation of the misinformation was provided (i.e., participants were given no explanation as to why the headline might be true or false).

social corrections provided by peers were equally effective in limiting misperceptions, which suggests that human connection is not necessary for the uptake of corrections.

Nonetheless, while automated content labelling offers advantages in terms of speed and scale of implementation, there are shortcomings with respect to consistency, reliability, and proven efficacy. First, classification algorithms are developed by both social media platforms and independent researchers, but the latter are disadvantaged by their dependence on publicly available data and the limited API access offered by some platforms (Freelon, 2018). Algorithms also usually remain secret while independent researchers typically open up their methods and results to scrutiny.

Second, automated labelling can be unreliable. Without human intelligence to review algorithmic judgments, there are significant risks of over-zealous and error-prone moderation (Banchik, 2020; Mühlhoff, 2019). Moreover, platforms typically fail to provide adequate information about their own efforts to measure the efficacy of technological interventions or to enable independent researchers to verify claims about efficacy. A review of COVID-19 interventions reported by platforms to the European Commission found that the application of technological interventions was highly inconsistent. For example, in some cases, generic warning labels were applied to accurate content while no warning labels were applied to misinformation content (Culloty et al., 2021). Such inconsistencies are concerning because public understanding of platform algorithms, content moderation procedures, and the application of content labels is limited.

Third, independent research on the efficacy of (automated) content labels is mixed. Some studies observed that content labels reduced intentions to share the labelled content (Mena, 2019) while other studies observed no effects on the perceived accuracy of the labelled content (Oeldorf-Hirsch et al., 2020). Aslett et al. (2022) found that news credibility labels failed to reduce misperceptions and had a limited effect on the quality of people's news diet. The content labels applied by social media platforms tend to be general warnings that caution people about the 'disputed' nature of claims or provide generic reminders to seek authoritative information. However, previous studies have found that general warnings are less effective than specific warnings (Clayton et al., 2020; Ecker et al., 2010).

Finally, scholars note that comprehension is likely to be influenced by the design of content labels and that much more research is needed to understand effective cross-cultural design for online platforms (Saltz et al., 2021). Such platforms, of course, have the capacity to answer these questions, but to date they have declined to share relevant information about who engages with their interventions, in which circumstances, and with what outcomes.

Recommendations for Policymakers and Tech Companies

In this paper, we have reviewed the evidence, knowledge gaps, and practical implications of four categories of misinformation interventions: boosting, nudging, debunking, and (automated) content labelling. We have focussed particularly on individual-level interventions with a grounding in psychological research. This categorisation is inevitably somewhat artificial. Interventions in each area overlap, and it is not always clear which interventions fall into which category.

Since 2016, misinformation research has experienced a massive boom, and many interventions have been shown to be effective at countering both the belief in and sharing of misinformation, particularly in laboratory settings. However, several open questions remain with respect to this growing body of research, which we translate into concrete recommendations for policymakers and tech companies.

First, much of the available research has focused heavily on Western Europe, North America, and Australia. Testing interventions designed for non-Western and non-English speaking countries remains complicated; for example, using online participant recruitment platforms such as Prolific Academic, Lucid or Respondi in developing countries may not yield sufficiently representative data, as samples are skewed towards highly educated city dwellers (Badrinathan, 2021). Alternative data collection methods such as phone surveys are much more difficult to conduct, and can be expensive. Importantly, what works in the West may not work elsewhere: interventions that were shown to be effective in Western countries did not yield the same results in countries such as India (Badrinathan, 2021; Guess et al., 2020) and Pakistan (Ali & Qazi, 2021). As Ali and Qazi (2021) note: the effectiveness of interventions "critically depends on how well their features and delivery are customised for the population of interest". Similarly, automated solutions may work well in English, but are typically not highly advanced in other languages. We encourage a rethinking of how to make conducting misinformation research outside of Western and developed countries more accessible and affordable, for example by investing in the representativeness of online samples, and making it easier for researchers to recruit participants through social media platforms.

Second, efficacy in the lab does not automatically translate to real-world effectiveness. DellaVigna and Linos (2022), for example, found that nudge interventions (e.g., communications such as letters or emails designed to increase vaccine uptake or reduce missed appointments) were about 6 times less effective (in terms of effect size) when implemented in the field compared to the lab. If this reduced effectiveness carries over to misinformation interventions (and there is reason to expect this; see Roozenbeek, van der Linden et al., 2022; Pennycook et al., 2021), it is important to ensure that interventions have robust effect sizes in lab studies before implementing them in the field. However, conducting field studies can be prohibitively expensive. Roozenbeek, van der Linden et al. (2022), for example, ran an advertisement campaign on YouTube to test the efficacy of a series of inoculation videos in an ecologically valid setting, with support from Google Jigsaw. Currently, the costs of such research are simply not affordable without collaborating with donors. We therefore recommend democratising efficacy testing of misinformation interventions in real-world settings, for example by providing researchers with free (or heavily discounted) ad credits and API access for social media platforms such as TikTok, Twitter, Facebook, and YouTube.

Third, to identify the causal impacts of misinformation interventions, studies typically focus on a single type of intervention. In this paper, we have discussed the evidence underlying how these interventions work in isolation. In practice, however, interventions co-

exist in a complex media-information environment. How these interventions work alongside each other is almost impossible to assess without cooperation from the platforms on which they are implemented, and there is a clear need for further multi-disciplinary and cross-platform collaborations.

Fourth, there is a great need for more open access data. In the wake of the COVID-19 pandemic, policymakers have become more forceful in calling on platforms to limit the spread of misinformation. The European Commission has acknowledged the need for greater access to data that will allow independent researchers to better understand the role platforms play in various domains, including mis- and disinformation. The Digital Services Act (European Commission, 2022) introduces obligations for platforms to make data available to "vetted researchers" (Art. 31) and this provision is echoed in the Guidance on Strengthening the Code (Art. 8.1). Some social media platforms have made access to their data available either through APIs (e.g., Twitter) or through curated services (e.g., Facebook's CrowdTangle). However, oftentimes data access remains problematic, and we call on platforms to improve transparency and data-sharing with researchers (Culloty et al., 2021). Greater access to platforms' data could provide a deeper understanding of the nature of misinformation and the impact of interventions in real-world settings while also addressing the issues noted above regarding geographical gaps in research and the prohibitive costs.

Fifth, in this paper we have focused primarily on solutions that are implemented at the level of the individual. As Chater and Loewenstein (2022) rightfully point out, doing so frames the misinformation problem in individual, not systemic terms. This risks drawing attention away from policies that seek to bring about systemic change. We argue that none of the interventions discussed in this paper, either individually or taken together, are enough to comprehensively address and counter misinformation. System-level solutions, such as requiring social media companies to give insight into their platforms' recommender algorithms (Alfano et al., 2021), are equally if not more important than individual-level interventions.

Sixth, outside of the regulatory sphere, companies such as Google and Twitter have also invested in the development and testing of misinformation interventions. Twitter, for example, ran a "prebunking" campaign to counter misinformation about election fraud ahead of the 2020 US presidential elections. Efficacy testing for this campaign was conducted internally, without input from independent researchers. To what extent the campaign contributed to a reduction in the spread of electoral misinformation is therefore not public knowledge. In addition, such experiments are often done without acquiring participants' informed consent (Grimmelmann, 2015). We therefore recommend more transparency about tech companies' efforts to measure the efficacy of interventions, and to enable independent researchers to verify their efficacy claims (Culloty et al., 2021). See Pasquetto et al. (2020) for a comprehensive review of how researchers and platforms may collaborate to counter misinformation.

Requiring platforms to ensure that independent experts vet misinformation interventions that they adopt, and their methods, data (privacy considerations permitting), and

conclusions are made public would be an important step forward. Ideally, however, independent experts would inform not only analyse the data collected by platforms, but also be instrumental in the design and implementation of efficacy testing. We therefore argue that all interventions must be designed and implemented with efficacy in mind. To achieve this, cooperation between academia and platforms is not enough: accountability is required.

Conclusion

Misinformation exists in an evolving environment that requires continuous monitoring. In this paper, we have discussed the evidence behind four categories of anti-misinformation interventions that take place at the individual level: boosting, nudging, debunking, and automated content labelling. Solutions that are effective today may work less well tomorrow, and actors who seek to spread misinformation deliberately also adapt to a changing environment. A comprehensive approach to tackling misinformation therefore involves 1) accountability and transparency on the part of tech companies and regulatory agencies in terms of how interventions are designed and tested; 2) collaboration between researchers (academic and non-academic) and tech companies not only in terms of data sharing and API access but the whole process of intervention design, efficacy testing, and implementation; 3) ensuring that creating and testing interventions in real-world environments and in non-Western settings becomes more affordable and accessible; 4) developing further insights into how different interventions work alongside each other in the real world; and 5) incorporating both individual-level (which we have discussed in this paper) as well as system-level approaches.

Funding Statement: J.R. has received research funding from the British Academy (#PF21\210010), Google Jigsaw, IRIS Coalition (UK Government, #SCH-00001-3391), the Economic and Social Research Council (ESRC, #ES/V011960/1), and JITSUVAX (EU Horizon 2020, #964728). J.S. and E.C. received funding from EU Horizon 2020 (Provenance, #825227), EU CEF (Ireland EDMO Hub, #2381686), and Broadcasting Authority of Ireland (CodeCheck 2019-2021).

Conflict of Interest: J.R. has been involved in the content development and testing of several interventions mentioned in this paper, such as the *Bad News*, *Harmony Square*, and *Go Viral!* games, as well as the inoculation videos at https://www.inoculation.science/.

References

Ahmed, W., Vidal-Alaball, J., Downing, J., & Seguí, F. L. (2020). COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. *Journal of Medical Internet Research, 22*(5), e19458. https://doi.org/10.2196/19458

Alaphilippe, A., Gizikis, A., Hanot, C., & Bontcheva, K. (2019). *Automated tackling of disinformation*. https://doi.org/10.2861/368879

Ali, A., & Qazi, I. A. (2021). Countering Misinformation on Social Media Through Educational Interventions: Evidence from a Randomized Experiment in Pakistan**.** *arXiv*. https://doi.org/10.48550/arXiv.2107.02775

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Altay, S., Berriche, M., & Acerbi, A. (2021). Misinformation on Misinformation: Conceptual and Methodological Challenges. *PsyArxiv Preprints.* https://doi.org/10.31234/osf.io/edqc8

Andı, S., & Akesson, J. (2021). Nudging Away False News: Evidence from a Social Norms Experiment. *Digital Journalism*, *9*(1), 106–125. https://doi.org/10.1080/21670811.2020.1847674

Arechar, A. A., Allen, J. N. L., Berinsky, A., Cole, R., Epstein, Z., Garimella, K., … Rand, D. G. (2022, February 11). Understanding and Reducing Online Misinformation Across 16 Countries on Six Continents. https://doi.org/10.31234/osf.io/a9frz

Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances, 8*(18). https://doi.org/10.1126/sciadv.abl3844

Au, C. H., Ho, K. K. W., & Chiu, D. K. W. (2021). The Role of Online Misinformation and Fake News in Ideological Polarization: Barriers, Catalysts, and Implications. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10133-9

Axelsson, C. W., Guath, M., & Nygren, T. (2021). Learning How to Separate Fake from Real News: Scalable Digital Tutorials Promoting Students' Civic Online Reasoning. *Future Internet, 13*(3), 60. https://doi.org/10.3390/fi13030060

Azzimonti, M., & Fernandes, M. (2018). *Social Media Networks, Fake News, and Polarization* (No. 24462). https://doi.org/10.3386/w24462

Badrinathan, S. (2021). Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. *American Political Science Review, 115*(4), 1325-1341. https://doi.org/10.1017/S0003055421000459

Banas, J. A., & Rains, S. A. (2010). A Meta-Analysis of Research on Inoculation Theory. *Communication Monographs*, *77*(3), 281–311. https://doi.org/10.1080/03637751003758193

Banchik, A. V. (2021). Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. *New Media & Society*, *23*(6), 1527–1544. https://doi.org/10.1177/1461444820912724

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data and Society*,

8(1). https://doi.org/10.1177/20539517211013868

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about Bad News: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3(1)*(2), 1–9. https://doi.org/https://doi.org/10.5334/joc.91

Beene, S., & Greer, K. (2021). A call to action for librarians: Countering conspiracy theories in the age of QAnon. *Journal of Academic Librarianship*, *47*(1), 102292. https://doi.org/10.1016/j.acalib.2020.102292

Benegal, S.D. and Scruggs, L.A. (2018). Correcting misinformation about climate change: the impact of partisanship in an experimental setting. *Climatic Change 148*(1–2), 61–80. https://doi.org/10.1007/s10584-018-2192-4

BMJ. (2021). *The BMJ will appeal after Facebook fails to act over "fact check" of investigation*. Www.Bmj.Com. https://www.bmj.com/company/newsroom/the-bmj-announces-appeal-after-facebook-fails-to-act-over-incompetent-fact-check-of-investigation/

Bode, L. & Vraga, E.K. (2018). See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication 33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Bontcheva, K., Posetti, J., Teyssou, D., Meyer, T., Gregory, S., Hanot, C., & Maynard, D. (2020). *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression*. https://unesdoc.unesco.org/ark:/48223/pf0000379015

boyd, d. (2018). *You Think You Want Media Literacy… Do You?* https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2

Braddock, K. (2019). Vaccinating Against Hate: Using Attitudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda. *Terrorism and Political Violence*. https://doi.org/10.1080/09546553.2019.1693370

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (1999). Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, *6*, 635–640. https://doi.org/10.3758/BF03212972

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(5). https://doi.org/10.1073/pnas.2020043118

Breakstone, J., Smith, M., Connors, P., Ortega, T., Kerr, D., & Wineburg, S. (2021). Lateral reading: College students learn to critically evaluate internet sources in an online course. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-56

Bulger, M., & Davison, P. (2018). The Promises, Challenges and Futures of Media Literacy. *Journal of Media Literacy Education*, *10*(1), 1–21. https://doi.org/10.23860/JMLE-2018-10-1-1

Chan, M. pui S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*. https://doi.org/10.1177/0956797617714579

Chater, N., & Loewenstein, G. (2022). The i-Frame and the s-Frame: How Focusing on Individual-Level Solutions Has Led Behavioral Public Policy Astray. *SSRN*. http://dx.doi.org/10.2139/ssrn.4046264

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Coleman, S. (2018). The elusiveness of political truth: From the conceit of objectivity to intersubjective judgement. *European Journal of Communication* 33(2), 157–171. https://doi.org/10.1177/0267323118760319

Compton, J. (2013). Inoculation Theory. In J. P. Dillard & L. Shen (Eds.), *The SAGE Handbook of Persuasion: Developments in Theory and Practice* (2nd ed., pp. 220–236). SAGE Publications. https://doi.org/10.4135/9781452218410

Compton, J., Van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602. https://doi.org/10.1111/spc3.12602

Conzola, V. C., & Wogalter, M. S. (2001). A communication-human information processing (C-HIP) approach to warning effectiveness in the workplace. *Journal of Risk Research*, *4*(4), 309–322. https://doi.org/10.1080/13669870110062712

Cook, J. (2021). Teaching students how to spot climate misinformation using a cartoon game. *Plus Lucis*, *3*, 13–16. https://crankyuncle.com/wp-content/uploads/2021/10/Cook_2021_Cranky_Uncle.pdf

Cook, J., Ecker, U. K. H., Trecek-King, M., Schade, G., Jeffers-Tracy, K., Fessmann, J., Kim, S. C., Kinkead, D., Orr, M., Vraga, E. K., Roberts, K., & McDowell, J. (2022). The Cranky Uncle game—Combining humor and gamification to build student resilience against climate misinformation. *Environmental Education Research*. https://doi.org/10.1080/13504622.2022.2085671

Cook, J., Ellerton, P., & Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, *13*(2). https://doi.org/10.1088/1748-9326/aaa49f

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, *12*(5), 1–21. https://doi.org/10.1371/journal.pone.0175799

Cookson, D., Jolley, D., Dempsey, R. C., & Povey, R. (2021). A social norms approach intervention to address misperceptions of anti-vaccine conspiracy beliefs amongst UK parents. *PLOS ONE*, *16*(11), e0258985. https://doi.org/10.1371/journal.pone.0258985

Craft, S., Ashley, S., & Maksl, A. (2017). News media literacy and conspiracy theory endorsement. *Communication and the Public*, *2*(4), 388–401. https://doi.org/10.1177/2057047317725539

Culloty, E., Park, K., Feenane, T., Papaevangelou, C., Conroy, A., & Suiter, J. (2021). *CovidCheck: Assessing the Implementation of EU Code of Practice on Disinformation*

in relation to COVID-19. https://doras.dcu.ie/26472/

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

DellaVigna, S., & Linos, E. (2022). RCTs to Scale: Comprehensive Evidence from Two Nudge Units. *Econometrica*, *90*(1), 81-116. https://doi.org/10.3982/ECTA18709

Duron, R., Limbach, B., & Waugh, W. (2006). Critical thinking framework for any discipline. *International Journal of Teaching and Learning in Higher Education*, *17*(2), 160-166.

Ecker, U. K. H., Sanderson, J., McIlhiney, P., Rowsell, J., Quekett, H., Brown, G., & Lewandowsky, S. (2022). Combining Refutations and Social Norms Increases Belief Change. *PsyArxiv Preprints*. https://doi.org/10.31234/osf.io/j9w8q

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*, 13-29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., & Antonio, L. M. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, *49*, 631–644. https://doi.org/10.3758/s13421-020-01129-y

Ecker, U. K. H., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), 41. https://doi.org/10.1186/s41235-020-00241-6

Ecker, U. K. H., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*, *111*(1), 36–54. https://doi.org/10.1111/bjop.12383

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory and Cognition*, *38*(8), 1087–1100. https://doi.org/10.3758/MC.38.8.1087

El Soufi, N., & See, B. H. (2019). Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence. *Studies in educational evaluation, 60*, 140-162. https://doi.org/10.1016/j.stueduc.2018.12.006

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-71

European Commission (2022). Digital Services Act Package. www.digital-strategy.ec.europa.eu. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Farkas, J., & Schou, J. (2020). *Post-Truth, Fake News and Democracy: Mapping the Politics of Falsehood*. Routledge.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the

sharing of false news. *Harvard Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-009

Fazio, L., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993–1002. https://doi.org/10.1037/xge0000098

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Freelon, D., & Wells, C. (2020). Disinformation as Political Communication. *Political Communication*, *37*(2), 145–156. https://doi.org/10.1080/10584609.2020.1723755

Funke, D., & Flamini, D. (2018). A guide to anti-misinformation actions around the world. *Poynter Institute for Media Studies*. https://www.poynter.org/ifcn/anti-misinformation-actions/

Garrett, R. K., & Bond, R. M. (2021). Conservatives' Susceptibility to Political Misperceptions. *Science Advances*, *7*(23), eabf1234. https://doi.org/10.1126/sciadv.abf1234

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The Effectiveness of Social Norms in Fighting Fake News on Social Media. *Journal of Management Information Systems*, *38*(1). https://doi.org/10.1080/07421222.2021.1870389

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Grady, R. H., Ditto, P. H., & Loftus, E. F. (2021). Nevertheless, partisanship persisted: fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications*, *6*(52). https://doi.org/10.1186/s41235-021-00315-z

Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism.* New York, NY: Columbia University Press. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226 FINAL.pdf

Grimmelmann, J. (2015). The Law and Ethics of Experiments on Social Media Users. *Colorado Technology Law Journal*, *13*(219), 219-271.

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*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guillory, J. J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility on the continued influence effect. *Journal of Applied Research in Memory and Cognition*, *2*(4), 201–209. https://doi.org/10.1016/j.jarmac.2013.10.001

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 178–206. https://doi.org/10.1162/tacl_a_00454

Hameleers, M., & van der Meer, T. G. L. A. (2019). Misinformation and Polarization in a

High-Choice Media Environment: How Effective Are Political Fact-Checkers? *Communication Research*, *47*(2), 227–250. https://doi.org/10.1177/0093650218819671

Hayes, C. (2006). 9/11: The roots of paranoia: Conspiracy theories and public mistrust. *The Nation*, *283*, 11–13.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *5*, 973–986. https://doi.org/10.1177/1745691617702496

Hobbs, R. (2021). *Media Literacy in Action: Questioning the Media*. Rowman & Littlefield Publishers.

Huber, C. R., & Kuncel, N. R., 2016. Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, *86*(2), pp.431-468. https://doi.org/10.3102/0034654315605917

Hughes, B., Braddock, K., Miller-Idriss, C., Goldberg, B., Criezis, M., Dashtgard, P., & White, K. (2021). Inoculating against Persuasion by Scientific Racism Propaganda: The Moderating Roles of Propaganda Form and Subtlety. *PsyArxiv Preprints*. https://doi.org/10.31235/osf.io/ecqn4

Human Rights Watch. (2018). *Germany: Flawed Social Media Law*. Www.Hrw.Org. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

International Press Institute. (2022). *Rush to pass 'fake news' laws during Covid-19 intensifying global media freedom challenges*. Www.Ipi.Media. https://ipi.media/rush-to-pass-fake-news-laws-during-covid-19-intensifying-global-media-freedom-challenges/

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*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*(8), 459–469. https://doi.org/10.1111/jasp.12453

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2019). Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist*, *65*(2), 371–388. https://doi.org/10.1177/0002764219869406

Kahne, J., & Bowyer, B. (2017). Educating for Democracy in a Partisan Age: Confronting the Challenges of Motivated Reasoning and Misinformation. *American Educational Research Journal*, *54*(1), 3–34. https://doi.org/10.3102/0002831216679817

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, *23*(5), 1301–1326. https://doi.org/10.1177/1461444820959296

Khan, I. (2021). *How Can States Effectively Regulate Social Media Platforms?* Oxford Business Law Blog. https://www.law.ox.ac.uk/business-law-blog/blog/2021/01/how-can-states-effectively-regulate-social-media-platforms

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Krause, N. M., Freiling, I., Beets, B., & Brossard, D. (2020). Fact-checking as risk

communication: the multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research*, *23*(7–8), 1052–1059. https://doi.org/10.1080/13669877.2020.1756385

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, N. M. (2018). Fake news, phishing, and fraud: a call for research on digital media literacy education beyond the classroom. *Communication Education*, *67*(4), 460–466. https://doi.org/10.1080/03634523.2018.1503313

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., … Zaragoza, M. S. (2020). *The Debunking Handbook 2020*. https://doi.org/10.17910/b7.1182

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "Post-Truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lewandowsky, S., & Yesilada, M. (2021). Inoculating against the spread of Islamophobic and radical-Islamist disinformation. *Cognitive Research: Principles and Implications*, *6*(57). https://doi.org/10.1186/s41235-021-00323-z

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01056-1

Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Boosting people's ability to detect microtargeted advertising. *Scientific Reports*, *11*(15541). https://doi.org/10.1038/s41598-021-94796-z

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*, *58*, 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964

Maertens, R., Anseel, F., & van der Linden, S. (2020). Combatting climate change misinformation: longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology*, *70*(101455). https://doi.org/10.1016/j.jenvp.2020.101455

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315

Margolin, D. B., Hannak, A., & Weber, I. (2017). Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication*, *35*(2), 196–219.

https://doi.org/10.1080/10584609.2017.1334018
McGuire, W. J. (1961). The Effectiveness of Supportive and Refutational Defenses in Immunizing and Restoring Beliefs Against Persuasion. *Sociometry*, *24*(2), 184. https://doi.org/10.2307/2786067
McGuire, W. J., & Papageorgis, D. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *Journal of Abnormal and Social Psychology*, *63*, 326–332. https://doi.org/10.1037/h0048344
Mena, P. (2019). Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet*, *12*(2), 165–183. https://doi.org/10.1002/poi3.214
Modirrousta-Galian, A., & Higham, P. A. (2022). How Effective are Gamified Fake News Interventions? Reanalyzing Existing Research with Signal Detection Theory. *PsyArxiv Preprints*. https://www.doi.org/10.31234/osf.io/4bgkd
Moore, T. (2014). Wittgenstein, Williams and the terminologies of higher education: A case study of the term 'critical'. *Journal of Academic Language & Learning*, 8(1), A95-A108. https://journal.aall.org.au/index.php/jall/article/view/314
Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021). Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445642
Mühlhoff, R. (2019). Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning. *New Media & Society*, 146144481988533. https://doi.org/10.1177/1461444819885334
Nuñez, F. (2020). Disinformation Legislation and Freedom of Expression. *UC Irvine Law Review*, *10*(2). https://scholarship.law.uci.edu/ucilr/vol10/iss2/10
Nygren, T., & Guath, M. (2021). Students Evaluating and Corroborating Digital News. *Scandinavian Journal of Educational Research*. https://doi.org/10.1080/00313831.2021.1897876
Nyhan, B. (2017). *Why the Fact-Checking at Facebook Needs to Be Checked*. Www.Nytimes.Com. https://www.nytimes.com/2017/10/23/upshot/why-the-fact-checking-at-facebook-needs-to-be-checked.html
Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2
Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The Ineffectiveness of Fact-Checking Labels on News Memes and Articles. *Mass Communication and Society*, *23*(5), 682–704. https://doi.org/10.1080/15205436.2020.1733613
Panizza, F., Ronzani, P., Martini, C., Mattavelli, S., Morisseau, T., & Motterlini, M. (2022). Lateral reading and monetary incentives to spot disinformation about science. *Scientific Reports*, *12*(1), 5678. https://doi.org/10.1038/s41598-022-09168-y
Pasquetto, I., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond,

R. M., Bozarth, L. C., Budak, C., Ecker, U. K. H., Fazio, L. K., Ferrara, E., Flanagin, A. J., Flammini, A., Freelon, D., Grinberg, N., Hertwig, R., Jamieson, K. H., Joseph, K., Jones, J. J. . . .Yang, K. C. (2020). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-49

Paynter, J., Luskin-Saxby, S., Keen, D., Fordyce, K., Frost, G., Imms, C., Miller, S., Trembath, D., Tucker, M., & Ecker, U. (2019). Evaluation of a template for countering misinformation: Real-world Autism treatment myth debunking. *PloS One*, *14*(1), e0210746. https://doi.org/10.1371/journal.pone.0210746

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation**.** *Nature Communications*, *13*(2333). https://doi.org/10.1038/s41467-022-30073-5

Petranová, D., Hossová, M. and Velický, P., (2017). Current development trends of media literacy in European Union countries. *Communication Today*, *8*(1), 52-64.

Piltch-Loeb, R., Su, M., Testa, M., Goldberg, B., Braddock, K., Miller-Idriss, C., Maturo, V., & Savoia, E. (2022). Testing the Efficacy of Attitudinal Inoculation Videos to Enhance COVID-19 Vaccine Acceptance: A Quasi-Experimental Intervention Trial. *JMIR Public Health and Surveillance*, *8*(6). https://doi.org/10.2196/34615

Potter, W. J., & Thai, C. (2016). Conceptual challenges in designing measures for media literacy studies. *International Journal of Media and Information Literacy*, (1-1), 27-42.

Pretus, C., Van Bavel, J. J., Brady, W. J., Harris, E. A., Vilarroya, O., & Servin, C. (2021). The role of political devotion in sharing partisan misinformation. *PsyArxiv Preprints*. https://doi.org/10.31234/osf.io/7k9gx

Quiring, O. *et al.* (2021). Constructive Skepticism, Dysfunctional Cynicism? Skepticism and Cynicism Differently Determine Generalized Media Trust. *International Journal of Communication 15*(22), 3497–3518.

Rathje, S., Van Bavel, J. J., Roozenbeek, J., & van der Linden, S. (2022a). Accuracy and Social Motivations Shape Judgements of (Mis)Information. *PsyArxiv Preprints*. https://doi.org/10.31234/osf.io/hkqyv

Rathje, S., Roozenbeek, J., Traberg, C. S., Van Bavel, J. J., & van der Linden, S. (2022b). Letter to the Editors of Psychological Science: Meta-Analysis Reveals that Accuracy Nudges Have Little to No Effect for US Conservatives: Regarding Pennycook et al.

(2020). *Psychological Science*. https://doi.org/10.25384/SAGE.12594110.v2

Reddy, P., Sharma, B., & Chaudhary, K. (2020). Digital Literacy: A Review of Literature. *International Journal of Technoethics, 11*(2), 65-94. https://doi.org/10.4018/IJT.20200701.oa1

Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy nudges? A pre-registered direct replication of Pennycook et al. (2020). *Psychological Science, 32*(7), 1–10. https://doi.org/10.1177/09567976211024535

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation. *Educational and Psychological Measurement*, *81*(2), 340–362. https://doi.org/10.1177/0013164420940378

Roozenbeek, J., Maertens, R., Herzog, S., Geers, M., Kurvers, R., Sultan, M., & van der Linden, S. (2022). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making 17*(3), 547-573.

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science 9*(211719). https://doi.org/10.1098/rsos.211719

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. https://doi.org/10.1080/13669877.2018.1443491

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Humanities and Social Sciences Communications*, *5*(65), 1–10. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., & van der Linden, S. (2020). Breaking Harmony Square: A game that "inoculates" against political misinformation. *The Harvard Kennedy School (HKS) Misinformation Review*, *1*(8). https://doi.org/10.37016/mr-2020-47

Roozenbeek, J., & van der Linden, S. (2022). How to Combat Health Misinformation: A Psychological Approach. *American Journal of Health Promotion*, *36*(3), 569–575. https://doi.org/10.1177/08901171211070958

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8*(34). https://doi.org/10.1126/sciadv.abo6254

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

Saleh, N., Roozenbeek, J., Makki, F., McClanahan, W., & van der Linden, S. (2021). Active inoculation boosts attitudinal resistance against extremist persuasion techniques – A novel approach towards the prevention of violent extremism. *Behavioural Public Policy*, 1–24. https://doi.org/10.1017/bpp.2020.60

Saltz, E., Barari, S., Leibowicz, C., & Wardle, C. (2021). Encounters with visual misinformation and labels across platforms: An interview and diary study to inform ecosystem approaches to misinformation interventions. *Extended Abstracts of the 2021*

CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery. https://doi.org/10.1145/3411763.3451807

Schmid-Petri, H., & Bürger, M. (2021). The effect of misinformation and inoculation: Replication of an experiment on the effect of false experts in the context of climate change communication. *Public Understanding of Science*, *31*(2), 152–167. https://doi.org/10.1177/09636625211024550

Swire-Thompson, B., Cook, J., Butler, L. H., Sanderson, J. A., Lewandowsky, S., & Ecker, U. K. H. (2021). Correction format has a limited role when debunking misinformation. *Cognitive Research: Principles and Implications*, *6*(83), https://doi.org/10.1186/s41235-021-00346-6

Swire-Thompson, B., Miklaucic, N., Wihbey, J., Lazer, D., & DeGutis, J. (2022). Backfire effects after correcting misinformation are strongly associated with reliability. *Journal of Experimental Psychology: General*. https://doi.org/10.31234/osf.io/e3pvx

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, *9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science*, *4*(3), 160802. https://doi.org/10.1098/rsos.160802

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News." *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2021). A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12551

Thorne, J., & Vlachos, A. (2018). Automated Fact Checking: Task Formulations, Methods and Future Directions. *Proceedings of the 27th International Conference on Computational Linguistics*, 3346–3359. https://aclanthology.org/C18-1283

Traberg, C. S. (2022). Misinformation: broaden definition to curb its societal influence. *Nature*, *606*, 653. https://doi.org/10.1038/d41586-022-01700-4

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *The ANNALS of the American Academy of Political and Social Science*. https://doi.org/10.1177/00027162221087936

Trammell, N. W., & Valdes, L. A. (1992). Persistance of negative priming: Steady state or decay? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 565–576.

Tully, M., Vraga, E. K., & Bode, L. (2020). Designing and Testing News Literacy Messages for Social Media. *Mass Communication and Society*, *23*(1), 22–46. https://doi.org/10.1080/15205436.2019.1604970

Ulbricht, L., & Yeung, K. (2022). Algorithmic regulation: A maturing concept for investigating regulation of and through algorithms. *Regulation & Governance*, *16*(3), 3–22. https://doi.org/10.1111/rego.12437

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political Psychology in the Digital (mis)Information age: A Model of News Belief and

Sharing. *Social Issues and Policy Review*, *15*(1), 84–113. https://doi.org/10.1111/sipr.12077

van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*. https://doi.org/10.1038/s41591-022-01713-6

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the Public against Misinformation about Climate Change. *Global Challenges*, *1*(2), 1600008. https://doi.org/10.1002/gch2.201600008

van der Linden, S., Roozenbeek, J., Maertens, R., Basol, M., Kácha, O., Rathje, S., & Traberg, C. S. (2021). How can psychological science help counter the spread of fake news? *Spanish Journal of Psychology*, *24*, 1–9. https://doi.org/10.1017/SJP.2021.23

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vraga, E. K., & Bode, L. (2020). Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation. *Political Communication*, *37*(1), 136–144. https://doi.org/10.1080/10584609.2020.1716500

Vraga, E. K., & Bode, L. (2017). Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication*, *39*(5), 621-645. https://doi.org/10.1177/1075547017731776

Vraga, E. K., Tully, M., & Bode, L. (2020). Empowering Users to Respond to Misinformation about Covid-19. *Media and Communication*, *8*(2). https://doi.org/10.17645/mac.v8i2.3200

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Walter, N., & Tukachinsky, R. (2020). A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It? *Communication Research*, *47*(2), 155–177. https://doi.org/10.1177/0093650219854600

Williams, M. N., & Bond, C. M. C. (2020). A preregistered replication of "Inoculating the public against misinformation about climate change." *Journal of Environmental Psychology*, *70*, 101456. https://doi.org/10.1016/j.jenvp.2020.101456

Wineburg, S., Breakstone, J., McGrew, S., Smith, M., & Ortega, T. (2022). Lateral reading on the open Internet: A district-wide field study in high school government classes. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000740a

Wong, N. C., & Harrison, K. J. (2014). Nuances in Inoculation: Protecting positive attitudes towards the HPV vaccine and the practice of vaccinating children. *Journal of Women's Health Issues & Care*, *3*(6). https://doi.org/10.4172/2325-9795.1000170

Wood, T., & Porter, E. (2019). The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Zerback, T., Töpfl, F., & Knöpfle, M. (2021). The disconcerting potential of online

disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society*, *23*(5), 1080–1093. https://doi.org/10.1177/1461444820908530

Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLOS ONE*, *12*(7), 1–27. https://doi.org/10.1371/journal.pone.0181821

Zollo, F., Novak, P. K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Emotional Dynamics in the Age of Misinformation. *PLOS ONE*, *10*(9), 1–22. https://doi.org/10.1371/journal.pone.0138740