7

Interventions to Combat Misinformation

7.1 INTRODUCTION

In the previous chapter, we discussed the legislative side of countering misinformation. But governments (and other regulatory entities) don't just have the law at their disposal: they (and nongovernment entities) can also leverage insights from psychology and behavioral science to reduce the spread of and susceptibility to misinformation. In this chapter we review the evidence behind the anti-misinformation interventions that have been designed and tested since misinformation research exploded in popularity around 2016. For this chapter, we draw on a review paper Jon wrote together with two colleagues, Eileen Culloty and Jane Suiter from Dublin City University (Roozenbeek, Culloty, et al., 2023). However, that paper is more expansive and detailed than this chapter can be for brevity reasons, and contains several recommendations for policymakers and tech companies that we're unable to cover here. Also, several other reviews have been published in recent years that employ somewhat different categorizations, discuss different publications, and arrive at different conclusions from us. These are worth having a look at, especially the very thorough review by Pica Johansson and colleagues (Bergsma & Carney, 2008; Czerniak et al., 2023; Y. Green et al., 2023; Gwiaździński et al., 2023; Hartwig et al., 2023; Janmohamed et al., 2021; Jeong et al., 2012; Johansson et al., 2022; Kozyreva et al., 2022; O'Mahony et al., 2023; Saltz et al., 2021; Traberg et al., 2022; Vahedi et al., 2018; van der Linden, 2022; Whitehead et al., 2023; Ziemer & Rothmund, 2022).

Anastasia Kozyreva and colleagues (2022) also put together a toolbox of interventions to counter misinformation. Their website (https://interventionstoolbox.mpib-berlin.mpg.de/) contains a lot of useful and practical information about the efficacy and applicability of many of the interventions we discuss in this chapter.[1] Finally, our research program has focused heavily on developing

[1] A research project sponsored by the US Social Science Research Council, led by Lisa Fazio, David Rand, and Stephan Lewandowsky, aims to compare the efficacy of these interventions in a single study (Social Science Research Council, 2022).

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

and testing a series of anti-misinformation interventions; we'll briefly mention this work here where relevant, but discuss it in much more detail in the next chapter.

7.2 TYPES OF MISINFORMATION INTERVENTIONS

There are two categories of intervention that seek to counter misinformation: *individual-level* and *system-level* interventions (Chater & Loewenstein, 2022; Kozyreva et al., 2020; Roozenbeek, Culloty, et al., 2023). System-level interventions tackle the *supply side* of misinformation, and include not only legislation (which we discussed in the previous chapter) but also changes to recommender algorithms (Guess et al., 2023a), addressing tech companies' business models, and political measures such as reducing polarization. Individual-level interventions target either people's behavior (usually what kinds of information they share with others on social networks) or susceptibility to misinformation (e.g., by reducing the likelihood of falling for misinformation). We've broken up individual-level interventions into four categories, using a modified categorization scheme originally developed by Anastasia Kozyreva and colleagues (2020) and previously used in the above-mentioned review paper (Roozenbeek, Culloty, et al., 2023): *boosting* skills or competences (media/digital literacy, critical thinking, and prebunking); *nudging* people by making changes to social media platforms' choice architecture; *debunking* misinformation through fact-checking; and (automated) *content labeling*. Figure 7.1, taken from Roozenbeek et al. (2023), shows an overview of the various system-level and individual-level interventions that are available to counter misinformation.

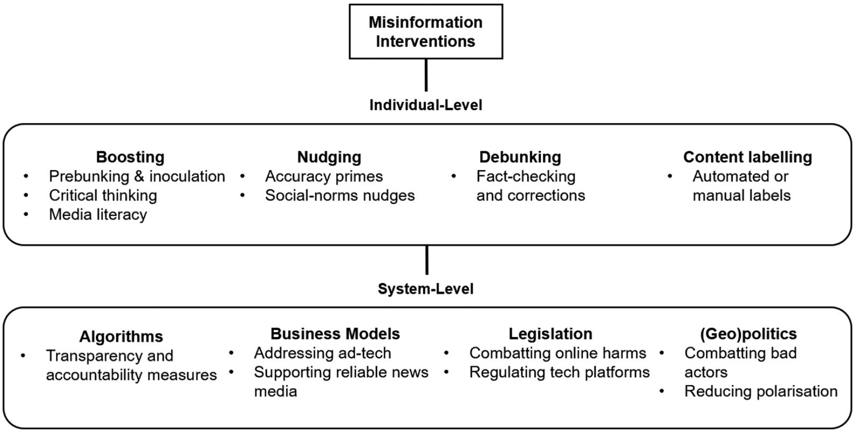

FIGURE 7.1 System-level and individual-level misinformation interventions.
Source: Taken from Roozenbeek, Culloty, & Suiter (2023).

7.3 BOOSTING

Boosting interventions are competence-focused, in the sense that they seek to "improve people's competence to make their own choices" (Hertwig & Grüne-Yanoff, 2017, p. 974).[2] Boosts are always voluntary and don't require making changes to people's choice environments. This makes them different from nudges, which are behavior-focused and involve making changes to how people interact with the information environment they encounter (more on this later). The benefit of boosting interventions is that they're noninvasive and are unlikely to pose substantial ethical challenges. For example, taking a media literacy class is unlikely to infringe on one's right to free expression. The downside of boosts is that they require people to decide to participate: if someone doesn't *want* to learn, they won't, and the boost won't be very beneficial to them. We'll discuss three types of boosting intervention: media and information literacy, critical thinking, and prebunking.

7.3.1 Media and Information Literacy

Media and Information Literacy (MIL) is an umbrella term that encompasses media literacy, information literacy, news literacy, and digital literacy (Carlsson, 2019; UNESCO, 2021). Media literacy focuses especially on young people, and covers not only trainings on how to spot mis- and disinformation, but also information on how to detect sponsored advertising, bias awareness, and empowering people to participate in media content production (Potter, 2020). Information literacy emphasizes competences such as finding and evaluating reliable information sources. News literacy focuses on content production and revolves around teaching people about how news and other kinds of content are produced (Tully et al., 2020). Finally, digital literacy interventions foster the skills required to navigate digital environments (Reddy et al., 2022).

Within the context of countering misinformation, MIL interventions are often designed for use in classrooms and other educational settings. To give a few examples, researchers at the University of Uppsala run the News Evaluator Project (Nyhetsvärderaren, https://nyhetsvarderaren.se/in-english), which is a series of instructional materials to help boost competences such as source criticism and civic online reasoning. These materials were extensively evaluated by the research team that developed them (led by our colleagues Thomas Nygren and Carl-Anton Axelsson), and are used in Swedish MIL curricula around the country (Axelsson et al., 2021; Nygren, 2019; Nygren et al., 2019; Nygren & Guath, 2019, 2021). Another high-profile example is the Civic Online Reasoning program (https://cor.stanford.edu/) led by Joel Breakstone

[2] A team of researchers led by Ralph Hertwig and Stefan Herzog put together a cool website about the science of boosting: www.scienceofboosting.org/.

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

and Sam Wineburg at Stanford University. This program includes a wide range of free educational materials and even full-fledged curricula for teaching skills such as critical ignoring, lateral reading, click restraint, and source critique. This research team has also conducted a wide range of evaluations of the effectiveness of various curricula, and shown that they are broadly effective at fostering key competences for navigating digital environments (Breakstone et al., 2021; McGrew et al., 2017, 2019; Wineburg et al., 2022).

A downside of MIL interventions that take place in classrooms is their scalability: most adults don't go to school and may be unlikely to voluntarily take a MIL class when offered. This means that the majority of the population won't immediately benefit from these kinds of intervention (Lee, 2018). Researchers have therefore looked for ways to deploy MIL interventions in online environments. For example, Folco Panizza and colleagues (2022) designed a social media pop-up that explains how to use lateral reading techniques (i.e., evaluating the credibility of a source by looking for additional information through search engines or other sources). They found that displaying the pop-up significantly increased the use of lateral reading strategies. Andrew Guess and colleagues (2020) tested whether reading a set of media literacy tips (a short, informative infographic with tips such as "check the credibility of the source") could help people better distinguish true from false information in a large study conducted in the United States and India. They found that the tips were highly effective at improving people's ability to spot false news content in both countries. However, in India, the intervention was only effective for a sample of highly educated individuals, and not for a sample of people from a mostly rural area.

Although MIL interventions are often effective at achieving their goals (boosting relevant competences), the research into their efficacy continues to suffer from several limitations. Most importantly, there is a distinct lack of research from non-Western countries, and what has been published doesn't always show encouraging results. For instance, both Sumitra Badrinathan (2021) and Ayesha Ali and Ihsan Ayyub Qazi (2021) found that educational interventions were mostly ineffective for rural participants in India and Pakistan, respectively. This shows that we lack a decent understanding of how to make interventions work in settings where especially Western researchers don't tend to go (Ghai, 2021, 2022). There is, however, also some cause for optimism: in a field experiment with about 9,000 participants in Kenya, Susan Athey and colleagues (2023) found that a five-day text message educational course was effective at reducing intentions to share misinformation. The treatment that was designed to counter emotion-based manipulation techniques was particularly effective (more so than the treatment that targeted reasoning-based techniques or a combination of both).

There are also conceptual challenges. MIL interventions research uses a wide range of measures and assessment methods, which makes it difficult to

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

compare different studies and interventions to one another (Potter & Thai, 2016). In addition, it's unclear if all types of MIL interventions are effective. A study by Mo Jones-Jang and colleagues (2019) showed that while information literacy was effective at boosting people's ability to identify misinformation, media, news, and digital literacy interventions were not; this suggests that more research is needed to probe what exactly makes MIL interventions effective. Finally, the efficacy of especially long-running MIL interventions (such as school curricula) is very difficult to assess. Longitudinal studies are hard and expensive to conduct, which makes optimizing interventions very complicated (Bulger & Davison, 2018).

7.3.2 Critical Thinking

Interventions aimed at boosting critical thinking typically intend to build skills related to people's ability to assess arguments, question underlying assumptions, and evaluate the quality of information (Duron et al., 2006). Critical thinking is related to media and information literacy, but not exactly the same: critical thinking may not be a domain-specific skill, but may instead be a transferable skill that can be applied in a variety of issue domains, not only misinformation (Axelsson et al., 2021; Moore, 2014).

In general, it appears that educational interventions are effective at teaching critical thinking. A meta-analysis about critical thinking in US college settings by Cristopher Huber and Nathan Kuncel (2016) concluded that "critical thinking skills and dispositions improve substantially over a normal college experience." However, the authors also note the following (conclusion section):

> Although the set of specific skills measured by critical thinking tests is important, spending more time on them involves trade-offs with other important skills. The evidence suggests that basic competencies such as reading and mathematics are more amenable to improvement beyond the gains currently observed, and the need is arguably more desperate.

In other words, curricula and other interventions that seek to boost critical thinking skills in specific domains are less effective than the overall effect of going to college. College students become better at critical thinking overall as they progress through their degree, whereas individual interventions don't appear to work very well. Still, within the domain of misinformation, Lauren Lutzke and colleagues (2019) found that critical thinking guidelines did boost people's ability to evaluate the credibility of misinformation (and non-misinformation) about climate change on Facebook. However, compared to MIL interventions and prebunking (see below), the research on critical thinking is limited: a systematic review by Paul Machete and Marita Turpin (2020) only identified three studies that directly dealt with critical thinking as a way to identify misinformation.

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

Other researchers have noted that many studies on critical thinking suffer from methodological shortcomings (El Soufi & See, 2019; Todd & O'Brien, 2016), which renders much of the evidence collected so far inconclusive.

Conceptually, scholars such as danah boyd (2018) have argued that the very notion of "critical thinking" can be weaponized by malicious actors seeking to confuse and obfuscate rather than clarify. They mention the example of RT (formerly Russia Today), a Kremlin-funded news outlet whose motto is "Question More." RT asking its readers to think critically about the science behind climate change, boyd argues, is not a genuine effort to appraise evidence in a better or more accurate way, but rather a way to leverage the idea of critical thinking to sow doubt about a well-established field of science. Adopting a critical stance by default, according to boyd and others (Beene & Greer, 2021), is not helpful. However, others have argued that it's important to distinguish between constructive skepticism and dysfunctional cynicism (Quiring et al., 2021); how to achieve this in a reliable way through interventions, however, isn't entirely clear.

7.3.3 Prebunking

The term "prebunking" (preemptive debunking) is now widely used, but we have no idea where it came from. We thought it was John Cook (another misinformation researcher who started using the term around 2016, around the same time as Sander), but when we asked him about it, he didn't know who came up with it either. We then looked online to find out when "prebunking" was first used, but to little avail. Nonetheless, prebunking refers to any kind of intervention that is deployed *before* people are exposed to misinformation. Or, according to Urban Dictionary (a top scholarly resource): "to debunk lies you know are coming in advance." While you could argue that this would mean that media literacy curricula should fall under the "prebunking" banner, the term usually refers to short-acting interventions that can be deployed on social media or are easily accessible online (e.g., browser games: see below). Various approaches to prebunking exist. For instance, Nadia Brashier and colleagues (2021) showed their study participants a simple banner that read "this article was rated false by independent fact-checkers" before they saw a false headline. Li Qian Tay and colleagues (2021) provided participants with a more detailed refutation of the misinformation that they saw shortly after (pointing out why the information was false or misleading).[3]

The most common framework for prebunking misinformation is inoculation theory, which was originally conceptualized in the 1960s by William

[3] For those interested in more practical resources, together with Jigsaw and BBC Media Action, we (but mostly Trisha Harjani, then a research assistant at our lab in Cambridge) put together a practical guide to designing and testing prebunking interventions (Harjani et al., 2022).

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

McGuire and Demetrios Papageorgis (McGuire, 1964, 1970; McGuire & Papageorgis, 1961b, 1962; Papageorgis & McGuire, 1961),[4] and later refined by scholars such as Michael Pfau, Bobi Ivanov, Kimberly Parker, John Banas, and Josh Compton (Banas & Miller, 2013; Compton, 2013, 2020, 2021; Compton & Pfau, 2009; Ivanov et al., 2011, 2012, 2018, 2022; Parker et al., 2012, 2016; Pfau, 1995; Pfau & Burgoon, 1988; Richards & Banas, 2018). Inoculation theory is grounded in a biological metaphor: much like how a medical vaccine (usually) consists of a weakened dose of the real pathogen, which prompts the body to produce antibodies, the process of psychological "vaccination" (or inoculation) involves preemptively exposing people to a "weakened dose" of an unwanted persuasion attempt, which should then increase resistance against subsequent persuasion attempts. The persuasive message is "weakened" by adding two components: a warning of an impending attack on one's beliefs or attitudes (e.g., "warning: some people might be out to manipulate you"), and a preemptive refutation of the upcoming manipulation attempt: for example, by explaining why the information is false (Compton, 2013). By doing this, people are both warned that their beliefs might be under attack and provided with the cognitive tools to resist future attempts to manipulate them. A meta-analysis by John Banas and Stephen Rains (2010) found that inoculation interventions are generally effective at conferring resistance against unwanted persuasion.

Inoculation theory became of significant interest for misinformation researchers around 2017. Both John Cook, Ulrich Ecker, and Stephan Lewandowsky (2017) and Sander and his colleagues at Yale (van der Linden, Leiserowitz, et al., 2017) tested whether inoculation could be used to reduce susceptibility to misinformation about climate change, both with some success: after inoculation, participants in both studies had a more accurate perception of the scientific consensus, and the inoculation treatment managed to reduce the adverse impacts of exposure to misinformation (van der Linden, Maibach, et al., 2017).[5]

There are two important theoretical distinctions when it comes to inoculation interventions: *passive* versus *active* inoculation (McGuire & Papageorgis, 1961a; Traberg et al., 2022), and *issue-based* versus *technique-based* (or logic-based) inoculation (Cook et al., 2017; Roozenbeek, Traberg, et al., 2022). Passive inoculation interventions provide people with the counterarguments needed to resist unwanted persuasion, for example by reading a piece of text or watching a video. Active inoculations, on the other hand, involve actively generating your own counterarguments: such interventions ask their participants to think

[4] Pronounced "papa gorgeous," you can't change our minds.

[5] It's useful to note that a replication of John Cook's study (Schmid-Petri & Bürger, 2021) didn't find the same effects as the original, in the sense that there was no effect of the inoculation on the perceived scientific consensus on climate change. This may have been due to the fact that the sample used in the replication (from Germany) generally already had lower baseline belief in misinformation compared to participants in the original study. A similar effect occurred in a replication of Sander's study from 2017 (Williams & Bond, 2020).

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

about why a piece of information might be false or misleading, and come up with their own strategies for countering it: for example, by playing a game.

Issue-based inoculation interventions tackle a specific argument or false claim that you don't want people to fall for: for example, specific misleading information about climate change (Maertens et al., 2020; van der Linden, Leiserowitz, et al., 2017; Williams & Bond, 2020). Technique-based inoculation interventions tackle the underlying rhetorical strategies and manipulation techniques that are often used to mislead or misinform, such as logical fallacies (Roozenbeek, van der Linden, et al., 2022), fake experts (Cook et al., 2017), trolling (Lees et al., 2023), or astroturfing (Zerback et al., 2021). The advantage of technique-based interventions over issue-based ones is that improving people's ability to spot a misleading rhetorical strategy potentially applies to a wide range of content, whereas issue-based interventions can only be expected to be effective for the specific argument or claim that people were inoculated against. However, if you can predict with reasonable certainty what misleading claims people are likely to be exposed to in the near future (e.g., about election fraud around an important election), issue-based interventions may be most effective.

Passive inoculations were shown to successfully boost resistance against misinformation in a variety of issue domains, including extremism (Braddock, 2019), COVID-19 (Basol et al., 2021), astroturfing comments (Zerback et al., 2021), and vaccine misinformation (Jolley & Douglas, 2017). More recently, researchers have begun to explore the use of short inoculation videos, which can be useful because video is a popular format that can easily be rolled out on social media: for example, as advertisements. Inoculation videos have been successfully tested in the realms of Islamist and Islamophobic propaganda (Lewandowsky & Yesilada, 2021), extremist propaganda (Hughes et al., 2021), and vaccine misinformation (Piltch-Loeb et al., 2022). We (Jon and Sander) have also helped create a series of inoculation videos and rolled them out as ads on YouTube (Roozenbeek, van der Linden, et al., 2022), but we will discuss this study in more detail in the next chapter.

Active inoculation interventions tend to come in the form of online games (Shi, 2023). The aforementioned John Cook, in collaboration with creative agency Goodbeast, created Cranky Uncle (www.crankyuncle.com/), a hilarious game where players learn about fourteen different techniques of science denial (in the context of climate change), such as the promotion of fake experts, cherry-picking, and various logical fallacies (Cook, 2020). The game is free to play and can be played in a browser or downloaded as an app. A qualitative evaluation of the Cranky Uncle game showed promising results for the game's use as an educational tool both inside and outside of classroom settings (Cook et al., 2022). Another very interesting (and effective) active intervention is the Spot the Troll quiz, in which people learn how to identify online trolls and other types of disingenuous messaging. Jeff Lees and colleagues (2023) showed that taking the quiz significantly increased people's

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

ability to distinguish troll comments from genuine content. Finally, we have also worked on a series of inoculation games ourselves, including Bad News (www.getbadnews.com/), Harmony Square (www.harmonysquare.game/), Go Viral! (www.goviralgame.com/), and Cat Park (www.catpark.game/). We'll discuss how we tested these interventions in the next chapter.

In general, prebunking interventions are mostly effective at reducing susceptibility to misinformation. However, they also have several important downsides (some of which we will discuss in Chapter 8). Most importantly, like MIL interventions, most prebunking interventions are rather lengthy and rely on voluntary uptake, and people who don't want to learn about how to spot misinformation will not benefit from them. This problem can be circumvented somewhat by making the interventions as entertaining as possible: for example, by using humor (Compton, 2018; Cook et al., 2022; Vraga et al., 2019). That said, what one person finds funny doesn't necessarily appeal to others,[6] making it necessary to continuously work on developing more and more interventions that appeal to different preferences. It's also important to note that people have to trust the source of the intervention; some people might be distrustful of an inoculation video about how smoking is secretly good for you produced by a tobacco company, and you might argue that this would be pretty reasonable. Similarly, if the production of prebunking interventions is not done in a transparent and open manner, people might be distrustful about their ultimate purpose, and refuse to engage with them.

Another limitation is that even successful prebunking interventions don't always exclusively impact people's evaluations of false or misleading information, but can also impact people's evaluation of factual information (Hameleers, 2023). This "real news" effect is commonly observed across a range of studies and for many different types of misinformation intervention (K. Clayton et al., 2020; Guess et al., 2020; Hoes et al., 2023; Modirrousta-Galian & Higham, 2023), although this discussion requires some nuance: prebunking interventions don't necessarily make people more skeptical of *all* information, but rather make them more skeptical of information that they already feel ambiguous about even without an intervention; their evaluations of information that is obviously true are not affected (Modirrousta-Galian & Higham, 2023). We discuss this phenomenon in detail in the next chapter.

7.4 NUDGES

In their famous book *Nudge: Improving Decisions about Health, Wealth, and Happiness* (2008, p. 6), Richard Thaler and Cass Sunstein (whose work on echo chambers we discussed in Chapter 5) define a "nudge" as "any aspect of the choice architecture that alters people's behavior in a predictable way without

[6] Looking at you, *Big Bang Theory*.

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

forbidding any options or significantly changing their economic incentive." Nudges come in many forms. One well-known example is moving the sugary or fatty products in the supermarket aisle from eye-level to foot-level to reduce how much unhealthy food people buy. In the context of misinformation, the unifying factor of nudges is that they seek to address some kind of unwanted information behavior, usually the sharing of misinformation with others. The potential advantages of nudges are numerous: they're cheap, nonintrusive, easy to implement, and highly scalable. Some social media companies have started implementing nudges in their choice architectures. Twitter, for instance, has asked people if they're sure they want to share or retweet a link if they haven't clicked on it.

The underlying assumption of nudge-based interventions is that a substantial proportion of social media users not only share unwanted content but are also "nudgeable" (de Ridder et al., 2021): people cannot be nudged into doing something they don't want to do. Nudgeable social media users therefore don't share misinformation knowingly and willingly (i.e., because they believe the misinformation to be true and want to tell others about it), but instead do so because they are in "irrational" modes of thinking. If these conditions are met, subtle changes to social media users' choice environments should positively affect their news-sharing decisions. Gordon Pennycook and David Rand (2019, 2021), whose work we also discussed in Chapter 4, have proposed that people tend to share misinformation (by which they mean false news headlines or "fake news": see Chapter 1) because of a failure to be mindful of accuracy. This, they argue, can happen because social media environments can be distracting and shift our attention away from the importance of sharing only accurate content with others. Their proposed solution is therefore to nudge people's attention *toward* accuracy: subtly reminding them of the concept of accuracy should prevent them from sharing misinformation when they encounter it in their social media feed. It's worth noting that this only works if people don't believe the misinformation that they see to be accurate to begin with; interventions that shift attention to accuracy only impact the sharing of misinformation that is accidental, and not deliberate.

Pennycook and Rand have proposed a series of interventions which they have collectively dubbed "accuracy prompts" or "accuracy nudges" (Epstein et al., 2021; Pennycook & Rand, 2022). There are different types of accuracy prompt intervention, which vary in terms of their intensity, effect size, length, and underlying mechanisms, but have in common that they all intend to improve the quality of people's news sharing decisions. The simplest and most well-known type of accuracy prompt consists of asking people to evaluate the accuracy of a single, nonpolitical news headline (Pennycook et al., 2020, 2021). People are shown a random news headline and asked "to the best of your knowledge, is the above headline accurate? Yes/No." This "evaluation" treatment should then activate the concept of accuracy in people's minds,

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

which subsequently improves the quality of their sharing decisions: they become either less likely to share false news headlines, more likely to share true news headlines, or both. This approach was shown to work in the context of both COVID-19 misinformation (Pennycook et al., 2020) and political misinformation (Pennycook et al., 2021). In a major publication in *Nature* (2021), Gordon Pennycook and colleagues also conducted a field study where they deployed this type of intervention on Twitter. They built a Twitter bot that sent people (a sample of mostly conservative US Twitter users) a direct message on Twitter, asking them to evaluate the accuracy of a single headline. Even though not every recipient read the message, the authors were still able to show that this "nudge" boosted the quality of the sources that people shared: although the effects were very small, Twitter users who received the nudge appeared to share slightly fewer news articles from sources with low quality ratings from fact-checkers (such as Breitbart) and more news from high-quality sources such as the *New York Times* and CNN. In an internal meta-analysis of accuracy prompt interventions, Pennycook and Rand (2022) concluded that overall, accuracy prompts are a "replicable and generalizable" approach for reducing the spread of misinformation.

However, this type of accuracy nudge intervention has also been met with some criticism. Several replications, including one we conducted, showed either no effect of the intervention on people's sharing decisions or a very small one (Gavin et al., 2022; Pretus et al., 2021; Rasmussen et al., 2022; Roozenbeek, Freeman, et al., 2021). Furthermore, four out of fourteen studies included in the accuracy prompt meta-analysis (Pennycook & Rand, 2022) showed no significant main effect, and a further three showed small effects that were just about significant. Thus, while an overall effect of this type of accuracy nudge on news-sharing intentions does appear to exist, this effect may be so small that you need a very large sample to detect it. This may be partially due to the fact that this type of intervention is a behavioral prime, where rather than giving people explicit instructions about the intended outcome of the intervention (e.g., "don't share misinformation with others!"), the intervention subtly and implicitly reminds people of the importance of accuracy. Such behavioral primes are known to be very difficult to replicate (Chivers, 2019; Schimmack, 2020; Schimmack et al., 2017).

Nonetheless, this criticism is specific to the single-headline "accuracy nudge" intervention, and other types of nudge (or prompt) intervention are more robust. For instance, Lisa Fazio (2020) found that asking people to pause for a while to consider why a headline is true or false substantially reduced people's willingness to share false headlines, and Pennycook and Rand (2022) report robust effect sizes for other "accuracy prompt" interventions such as PSA videos and media literacy tips. In addition, social norms around the sharing of news and other content also appear to be a promising avenue for intervention: both Andı and Akesson (2021) and Gimpel and colleagues

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

(2021) found that emphasizing injunctive norms (what behavior most people approve or disapprove of: for example, "most people think sharing fake news is bad") could reduce the proportion of people who are willing to share false news with others. Ziv Epstein and colleagues (2021) put together an "accuracy prompt toolkit" that contains many different kinds of nudge intervention. Finally, in a field experiment conducted on Twitter, Matthew Katsaros and colleagues (2021) designed an intervention that prompted Twitter users about to post harmful or hateful content with an opportunity to pause and reconsider what they wrote. Promisingly, this intervention led to a 6 percent reduction in the posting of offensive tweets, compared to a control group. This highlights the real-world applicability of nudge interventions.

Although nudges are a highly promising approach to countering misinformation, they're also beset by several challenges. Most importantly, it's incredibly difficult to translate promising findings from lab studies to real-world settings. Stefano DellaVigna and Elizabeth Linos (2022) compared the effectiveness of nudge interventions (unrelated to misinformation) in lab studies and that of the same interventions when implemented in the field. They found that interventions were about four to six times less effective in the field than in the lab (although the effect remained detectable). This means that we can expect a substantial reduction in effect size when implementing nudge interventions in settings where people don't have to pay attention, which highlights the importance of starting out with robust effect sizes in lab studies. Furthermore, it appears that nudges become less effective the more people are exposed to them. A study by Shusaku Sasaki and colleagues (2021) found that the nudge effect wore off after only a few exposures, indicating that nudges may not retain their initial effectiveness (although we need more research to verify if this is the case on social media as well).

7.5 DEBUNKING AND FACT-CHECKING

Debunking involves correcting misinformation after exposure (Bode & Vraga, 2018).[7] It's one of the most popular approaches to tackling misinformation: websites such as Snopes, FullFact and StopFake have large numbers of followers and reach millions of people, and many tech companies have extensive fact-checking policies in place. Meta (formerly Facebook) runs a worldwide fact-checking program where it pays third-party fact-checkers to evaluate and label content that is posted on Facebook, WhatsApp, and Instagram.[8]

[7] Debunking and fact-checking are similar terms but not entirely synonymous: debunking always pertains to misinformation, whereas you could fact-check a story and rate it "true" (Roozenbeek, Culloty, et al., 2023).

[8] For full disclosure: Sander has been an adviser to Meta's fact-checking program. He can't use the Zuckerbucks he receives as compensation to buy anything other than more Zuckerbucks, but he's told that they go up in value over time.

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

Luckily, there appears to be scientific consensus about the effectiveness of debunking. A series of meta-analyses and review papers (Chan et al., 2017; Walter et al., 2020; Walter & Murphy, 2018; Walter & Tukachinsky, 2020) has shown that, overall, correcting misperceptions reduces belief in misperceptions. This sounds almost tautological, but it's an important finding: overall, people are willing to change their belief in a false claim when presented with corrective information. Research teams like the one run by Andreas Vlachos (Guo et al., 2022; Schlichtkrull et al., 2023; Thorne & Vlachos, 2018) have made great strides in improving automated fact-checking, which can potentially scale up fact-checking efforts on social media to a considerable degree. In 2020 a team of researchers led by Stephan Lewandowsky, John Cook, and Ulrich Ecker put together a practical guide to debunking, the *Debunking Handbook*, which explains the science of debunking and how to leverage it for effectively correcting misinformation online (Lewandowsky et al., 2020).

In the 2010s, concerns were raised about debunking accidentally provoking an *increased* belief in misperceptions (Nyhan & Reifler, 2010; C. Peter & Koch, 2015), a phenomenon which became known as the "backfire effect." For example, a study led by Mohsen Mosleh (2021) found that Twitter users who were corrected for posting false political news increased the subsequent sharing of toxic and low-quality content. However, recent review studies have shown that these backfire effects are extremely rare and not reliably observed (Ecker, Lewandowsky, et al., 2020; Swire-Thompson et al., 2020, 2022; Wood & Porter, 2019). The risks of debunking "side effects" are therefore considered to be very low.[9]

That said, like all misinformation interventions, debunking faces several challenges. First, debunking is not universally effective for everyone across all issue domains: correcting health misinformation appears to be easier than correcting misinformation about politics and marketing (Chan et al., 2017; Porter & Wood, 2021; Walter et al., 2020; Walter & Murphy, 2018). A recent meta-analysis on the effectiveness of debunking science-relevant (including health-related) misinformation found no overall significant effects (Chan & Albarracín, 2023), particularly for misinformation about politically polarizing topics. The content of the debunking message itself matters as well. For example, less detailed fact-checks appear to be less effective than more detailed ones (Ecker, O'Reilly, et al., 2020; Paynter et al., 2019). Another problem is source credibility: how trustworthy people perceive the source of a fact-check to be strongly affects how likely someone is to accept a correction. In other words, if you don't like the source of the fact-check, you likely won't believe it (Bode & Vraga, 2018, 2021; Ecker & Antonio, 2021; Guillory & Geraci, 2013; Vraga & Bode, 2017). A study led by Drew Margolin (2017) also showed that Twitter users were much more likely to accept a fact-check if it came from an account

[9] Still, like prebunking (as we'll see in the next chapter), every anti-misinformation intervention has benefits as well as potential side effects to consider.

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

that they followed than if it came from a stranger. Problematically, Michael Hameleers and Toni van der Meer (2019) found that people are much more likely to engage with corrections that are congruent with their prior (political) attitudes, and usually ignore those that contradict their prior beliefs.

More generally, the study on debunking on Facebook by Fabiana Zollo and colleagues (2017), which we discussed in Chapter 5, showed that fact-checks don't reach a lot of people who tend to consume conspiratorial content, possibly due to echo chamber formation. This means that it may be exceedingly difficult to provide effective corrections to social media users with the highest belief in misinformation; pessimistically speaking, many fact-checks may be preaching to the converted. Moreover, even if a correction successfully reduces a misbelief, it doesn't completely eliminate it: the "continued influence effect" (Ecker & Antonio, 2021; Lewandowsky et al., 2012; Walter & Tukachinsky, 2020) states that people continue to (partially) rely on misbeliefs and retrieve them from memory even after it has been successfully corrected. It's believed that this happens because information that was previously encoded in people's memory can continue to influence one's judgments, even if more recent information contradicts it (H. M. Johnson & Seifert, 1994).

Finally, as we discussed in Chapters 1 and 3, the most impactful misinformation is often not explicitly false, and judging whether something counts as misinformation can be highly subjective (Coleman, 2018). Because of this, fact-checking can be highly contentious and risks becoming politicized (Graves, 2016). This problem is compounded by the fact that some fact-checkers are dependent on large donors. Meta's third-party fact-checking program, for example, has been criticized for its lack of transparency (Sander denies any responsibility), and some have emphasized the risk of placing responsibility for what kinds of content are and aren't fact-checked in the hands of large corporations (BMJ, 2021; Nyhan, 2017).

7.6 (AUTOMATED) CONTENT LABELING

Because of the sheer volume of content uploaded to social media platforms, implementing misinformation interventions at scale is a daunting challenge. Some platforms therefore rely on automated methods to label and moderate content (Alaphilippe et al., 2019). Many of these methods rely on machine learning, and generally either classify content into categories (such as "misleading" or "false") or evaluate it against a database of known problematic sources or claims (Thorne & Vlachos, 2018). Content labels can come in the form of general or specific warnings (K. Clayton et al., 2020; Mena, 2019), fact-checks (Brashier et al., 2021), or news credibility labels (Aslett et al., 2022). Well-known examples of labels are NewsGuard, which rates the credibility of news sources, and Facebook's content labels, which provide context to what people are seeing on the platform (such as "satire page," "public official," and

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

“fan page”). Twitter now also has a feature called Community Notes, where users can collaboratively add context to potentially misleading tweets.

Vincent Conzola and Michael Wogalter (2001) note that for content labels to be effective, they must attract enough attention for the information to be noticed, be clear enough that people can understand the message that is conveyed, and motivate people enough to take the required action (e.g., not sharing something with others when they otherwise would have). And indeed, Tatiana Celadin and colleagues (2023) found that displaying the trustworthiness of a news source reduced people’s intentions to share false news posts. In a study on the effects of warning labels that were applied to Donald Trump’s tweets about the 2020 US presidential elections, Orestis Papakyriakopoulos and Ellen Goodman (2022) found that, while adding labels didn’t change Twitter users’ engagement with or sharing of Trump’s tweets overall, adding strong rebuttals and a degree of contextual overlap between the label and the tweets did reduce engagement. They conclude that the right kinds of labels may be a “plausible way to mitigate misinformation spread and toxic user engagement.”

However, not all content labels appear to be effective. In a large study also conducted on Twitter, Kevin Aslett and colleagues (2022) investigated if news credibility labels can impact the quality of people’s news diets and reduce belief in misinformation. They found that this was not the case: overall, the labels provoked very limited effects on news diet quality shortly after being introduced, and didn’t reduce misperceptions. Similarly, Anne Oeldorf-Hirsch and colleagues (2020) found that fact-checking labels had little to no effect on the perceived accuracy of news memes and articles. It’s also worth noting that automated detection methods are imperfect and can be unreliable; without humans to review judgments made by algorithms, there’s always a risk of error-prone moderation (Banchik, 2021).

7.7 SOME REFLECTIONS ON MISINFORMATION INTERVENTIONS

With some exceptions, individual-level misinformation interventions appear to be effective at what they set out to achieve, be it reducing misperceptions, increasing resilience to manipulation, or nudging people into changing their (self-reported) behavior. However, are these kinds of individual-level solutions really the way to go? In an influential paper, Nick Chater and George Loewenstein (2022) argue that an excessive focus on individual-level solutions to societal problems (including but not limited to tackling misinformation) has “led behavioral public policy astray.” They note that while some individual-level solutions are effective in the statistical sense, they at best combat the symptoms of larger underlying problems (Altay, 2022; Roozenbeek & Zollo, 2022). An overreliance on simple solutions and “quick fixes” may therefore draw necessary attention away from systemic problems, such as tech

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

companies' recommender algorithms showing people dodgy content because it's good for ad revenue.

Then again, it's also true that individual-level misinformation interventions are unlikely to cause harm: fact-checks, prebunks, and nudges don't violate people's rights, even if some people may find them annoying. System-level solutions, such as legislation that puts limits on what kinds of content people can post online, may be much more effective in terms of reducing exposure or sharing of misinformation, but also carry significant risks such as posing a threat to freedom of expression (see Chapter 6). For example, analyses find that misinformation about election fraud dropped significantly after Twitter's decision to ban Donald Trump (Dwoskin & Timberg, 2021), but this also prompted Trump to start his own social media platform (Truth Social), resulting in audiences becoming even more fragmented and potentially more insulated from factual information. Cass Sunstein (2023) also disagrees strongly with the notion that individual-level solutions (such as nudges) can "crowd out" more aggressive approaches, calling it "preposterous." He argues that there is no evidence that the implementation of individual-level solutions makes system-level reforms less likely to occur.

To illustrate this, let's revisit Figure 1.5 from Chapter 1, but this time let's also look at what interventions might be considered for each type of content (verifiable falsehoods, misinformation that is misleading but not outright false, true but ambiguous information, and uncontroversial facts). In Chapter 1 we noted some of the difficulties with defining "misinformation": verifiable falsehoods and uncontroversial facts don't cause too much trouble in this respect, but it can be tremendously difficult to distinguish between misinformation that isn't outright false and non-misinformation of ambiguous veracity (e.g., because relevant context is left out, and so on).

Looking at Figure 7.2, we see that the difficulties we described in Chapter 1 rear their heads again when deciding on the appropriate intervention(s) to counter misinformation. Again, verifiable falsehoods and uncontroversial facts are relatively unproblematic: falsehoods can be removed or downranked by social media platforms without too much controversy, and other solutions such as nudging, debunking, content labeling, media literacy, and prebunking are all potentially effective as well. Uncontroversial (or reliable) news content can be labeled as such, digital literacy programs can educate people about how to identify reliable sources, and platforms can boost the visibility of reliable and trustworthy content. Our problems start when we try to tackle misinformation that isn't entirely false: such content is difficult to remove or downrank (because there might be substantial disagreement over whether it should be labeled as "misinformation," thus potentially evoking controversy). Accuracy prompts (see above) are known to work less well for misinformation that is seen as more persuasive (Arechar et al., 2022), and debunking, literacy, labeling, and prebunking interventions can work in principle but also require

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

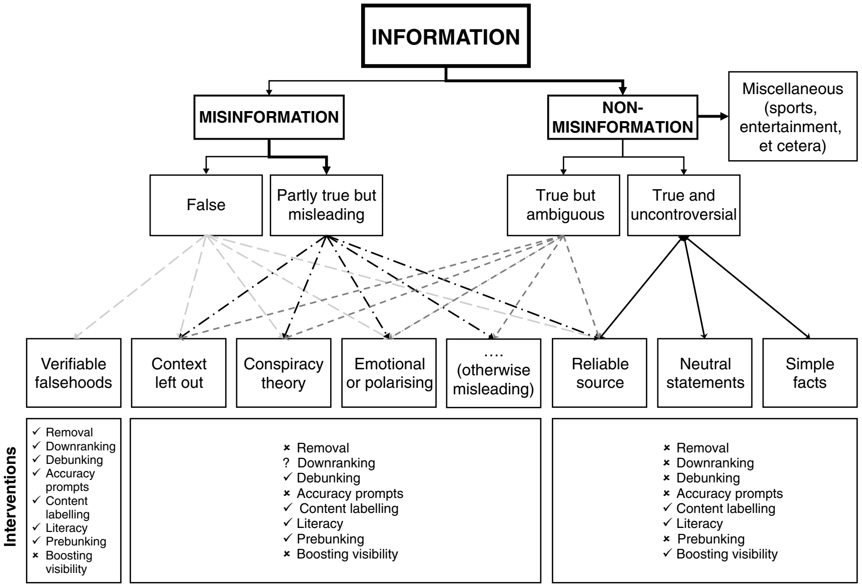

FIGURE 7.2 Flowchart for defining misinformation, with interventions. Crosses, ticks, and question marks indicate whether an intervention type is not suitable, suitable, or questionably useful for that type of (mis)information.

careful consideration: after all, you don't want to accidentally label something as misinformation that turns out to be true.

With all this in mind, the body of evidence that has been amassed over the last few years on the efficacy of misinformation interventions is growing rapidly, but continues to suffer from several shortcomings. Most importantly, as is a recurring theme throughout this book, there's a huge research gap for non-Western countries: we simply lack a lot of knowledge on what works and what doesn't in countries outside of the US and Western Europe. There's also not enough evidence on what works in the field; lab studies abound, but as we discussed above, it's not necessarily true that what works in a lab also works in real life (DellaVigna & Linos, 2022). This problem is compounded by the fact that researchers often lack access to social media data, and conducting experimental studies in social media settings is not only complicated but also extremely expensive (Roozenbeek & Zollo, 2022). There's a need to make access to data and funding more democratic and accessible, especially for researchers who don't work at rich, Western universities.

7.8 CONCLUSION

In this chapter, we've discussed the effectiveness of four categories of individual-level misinformation interventions: boosting (media and information literacy, critical thinking, and prebunking); nudging; debunking (and fact-checking);

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press

and (automated) content labeling. These interventions have one of three goals: to improve relevant skills such as spotting manipulation techniques, source criticism, or lateral reading (in the case of boosting interventions and some content labels); to change people's behavior, most commonly improving the quality of their sharing decisions (for nudges and most content labels); or to reduce misperceptions and misbeliefs (in the case of debunking). We've argued that at least some interventions from each category are effective at what they set out to achieve, and more and more evidence continues to be gathered that builds our knowledge of what works and what doesn't. However, there are still open questions with respect to the cross-cultural and real-world effectiveness of many interventions, including over time, and there are important issues to consider when it comes to whether focusing too much on "fixing" individual beliefs or behavior doesn't distract from implementing more systemic solutions. In the next and final chapter, we will discuss our own research program on misinformation interventions. We will talk about what worked and what went well, but also about the nuances and limitations to our work.

https://doi.org/10.1017/9781009214414.011 Published online by Cambridge University Press