# **ESSAYS ON RESISTANCE AGAINST PERSUASION**

Building, strengthening, and spreading attitudinal resistance through inoculation theory.



Melisa Basol Pembroke College

Social Decision-Making Research Lab Department of Psychology University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy, October 2022

## DECLARATION

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university of institution for any degree, diploma, or other qualification.

In accordance with the Department of Psychology guidelines, this thesis does not exceed 60,000 words and contains less than 150 figures.

Signed:\_\_\_\_\_

Date:

Melisa Basol, MPhil Cambridge

### **SUMMARY**

The prevalence of misinformation is a threat to science, society, and the democratic process. Current efforts are mostly reactive and consist of predominantly legislative, algorithmic, and educational interventions. However, growing psychological research emphasises the difficulty of catching up with and undoing the harms of manipulative content once it is out, calling for pre-emptive efforts that could stop harmful information from going viral in first place. Though the efficacy of inoculation theory, often regarded as the "grandfather theory of persuasion", has been demonstrated across varying contexts, little research exists on its efficacy against online misinformation. The aim of this doctoral research was to examine how inoculation theory may be used to combat misinformation. To do so, I sought to establish how attitudinal resistance to misinformation can be build, strengthened, and spread by designing and testing novel theory-driven interventions using

randomized experiments in both the lab and the field. Across several empirical studies, results consistently suggest that generalised and gamified inoculation treatments are effective in reducing the perceived reliability of misinformation, in boosting attitudinal certainty, and in decreasing people's willingness to share manipulative information.

More specifically, in Chapter 1, I test the efficacy of "Bad News" as an inoculation treatment against common manipulation strategies and found that the intervention significantly increases people's ability to spot misinformation techniques and boosts their level of confidence in their own (correct) judgements. These findings are further extended in Chapter 2, where I demonstrate the efficacy of a new gamified and generalised inoculation treatment within the context of end-to-end encrypted private messaging apps and extend the findings on attitude certainty by identifying it as a significant mediator for sharing intentions of misinformation- emphasising the crucial role of certainty when resisting. Additionally, Chapter 2 finds that inoculated individuals are significantly less likely to share content that includes manipulative content. Chapter 3 further replicates and builds on these findings by providing additional and longitudinal support for new gamified inoculation treatments across three different languages in the context of the COVID-19 pandemic. I evaluated a real-world intervention adopted by the UK government and World Health Organization, empirically demonstrating that it improves reliability assessments of misinformation, improves people's certainty in their ability to spot and resist misinformation, and reduces self-reported willingness to share misinformation with others in their social network. Chapter 3 also takes a critical look at the role of apprehensive versus motivational threat, one of the theoretical tenants of inoculation theory. In Chapter 4, I explore the effects of post-inoculation talk on the inoculated participants as well as those who vicariously receive second-order inoculation treatments through talk. These findings provide novel contributions to whether it has the potential to keep up with and outpace the speed and depth at which online misinformation travels. Specifically, content analyses provide novel insights into how and when postinoculation talk occurs and, more importantly, what it is about. Thus, this doctoral research makes novel use of post-inoculation talk by pivoting from intra-individual resistance to inter-individual resistance. By demonstrating the effectiveness of receiving vicarious inoculation treatments, this research contributes to the quest for psychological herd immunity against misinformation. In sum, this doctoral research sheds light on the antecedents that underpin the inoculation process and how resistance against misinformation can be build, strengthened, and ultimately spread from one individual to another.

### ACKNOWLEDGEMENTS

My time at Cambridge is marked by the freedom to contemplate, be curious, and grow as a person and scholar alike. Along the way, I got to be part of something I would not have dared to dream of. I am convinced that no one who succeeds does so alone – and in my case, much of this is thanks to the trust and support of individuals I wish to briefly acknowledge here.

Firstly, I am beyond grateful to my supervisor, Prof Sander van der Linden, whose mentoring always encouraged me to challenge the status-quo and give the unimaginable a try. Leading by example, he has instilled in me the importance of scientific rigour, our societal responsibility to use our work for good, and what I have come to admire the most, his relentless optimism that a better a world is possible. Our world has seen many challenges over the course of my doctoral degree and yet, Sander never stopped approaching them with an energetic tenacity and an unshakeable commitment to be part of the solution. Yet, increasing recognition and ego have never prevented him from giving me, and all his other students, the chance to shine. Instead, he built the Social Decision-Making Research Lab, a collaborative space of witty, energetic, and undeniably prolific scholars who have all, explicitly and

implicitly, contributed to my work. Lastly, I am eternally grateful for Sander's patient and kind guidance when I lost sight and doubted my abilities. I can only hope that I will do his mentoring justice one day.

Secondly, I am thankful for the Gates Cambridge Trust. Their faith in me has allowed me to take on this doctoral degree and has introduced me to a community of scholars dedicated to applying their work to pressing societal issues. Being a Gates Scholar substantially marked my experience at Cambridge and has shown me that believing one can better the world might be the most courageous thing one can do. I hope to reflect their principles from here on and, most importantly, am eternally grateful for the dear friends I have made by being a part of it all.

Lastly, I want to thank my mother. The fierce and brilliant woman who single-handedly managed to raise her children and, by sacrificing everything, build us a ladder to the stars.

### TABLE OF CONTENTS

### **1 INTRODUCTION** 14

**1.1 GENERAL INTRODUCTION14**SOCIETIES' STATE OF AFFAIRS14THE HISTORY OF MISINFORMATION AND ITS ADAPTATIONS TO THE DIGITAL AGE 16CURRENT EFFORTS TO COUNTER MISINFORMATION17ORIGINS OF INOCULATION THEORY 19

### **2 THE ROLE OF ATTITUDE CERTAINTY IN RESISTANCE TO PERSUASION 30**

#### 2.1 ABSTRACT 30 31 **2.2 INTRODUCTION** 2.2.1 PERSUASION RESEARCH ON ATTITUDE CERTAINTY 31 2.2.2 INOCULATION RESEARCH ON ATTITUDE CERTAINTY 32 PUSHING THE THEORETICAL BOUNDARIES: INOCULATING AGAINST MISINFORMATION 33 FINDING AN ANTIDOTE: GAMIFIED INOCULATION 34 SHORTCOMINGS AND REMAINING GAPS 35 2.3 METHODS 35 PARTICIPANTS AND PROCEDURE 35 MEASURES 37 37 2.4 RESULTS 2.4 DISCUSSION 40 THE CASE FOR THERAPEUTIC AND GENERALISED INOCULATIONS 41 SITUATING ATTITUDE CONFIDENCE IN THE INOCULATION PROCESS 41 LIMITATIONS AND FUTURE RESEARCH 42 **2.5 CONCLUSION** 42

### <u>3 THE ROLE OF ATTITUDE CERTAINTY IN ONLINE SHARING INTENTIONS 44</u>

### 3.1 ABSTRACT 44

### 3.2 INTRODUCTION 45

MISINFORMATION PROTECTED BY THE WALLS OF END-TO-END ENCRYPTIONS 45 THE PSYCHOLOGICAL DYNAMICS ON PRIVATE MESSAGING PLATFORMS 46 **INOCULATION THEORY – THEORETICAL BOUNDARIES AND EXTENSIONS 47** THE ROLE OF ATTITUDE CERTAINTY IN RESISTING PERSUASION 48 49 CARING IS SHARING IN PURSUIT OF A SOLUTION WITH WHATSAPP 50 **3.3 DEVELOPING JOIN THIS GROUP** 51 3.4 METHODS 52 SAMPLE 52 MEASURES 53 PROCEDURE 55 **3.5 RESULTS 58** 58 Reliability CONFIDENCE 60 SHARING 62 3.6 DISCUSSION 65 GAMIFIED INOCULATION ACROSS TOPICS, PLATFORMS, AND INDIVIDUALS 65 GENERATING, STRENGTHENING, AND SPREADING INOCULATION THROUGH ATTITUDE CERTAINTY 66 LIMITATIONS AND REMAINING QUESTIONS 67 SHIFTING TOWARDS THE SPREAD OF INOCULATION 69 **3.7 CONCLUSION** 69

# 4 TOWARD PSYCHOLOGICAL HERD IMMUNITY: CROSS-CULTURAL EVIDENCE FORTWO PREBUNKING INTERVENTIONS AGAINST COVID-19 MISINFORMATION71

| 4.1 ABSTRACT    | 72         |          |            |           |        |    |
|-----------------|------------|----------|------------|-----------|--------|----|
| 4.2 INTRODUCT   | ION        | 72       |            |           |        |    |
| REVISITING THE  | BASICS     | 73       |            |           |        |    |
| ESSENTIAL THEO  | RETICAL    | BUILDIN  | G BLOCKS   | 74        |        |    |
| WHY APPREHENS   | SION IS NO | OT THE S | OLUTION 2  | 75        |        |    |
| BRINGING MOTIV  | ATION BA   | АСК ІМТО | THREAT 2   | 75        |        |    |
| ATTITUDINAL IN  | OCULATIO   | )N DURIN | IG A GLOBA | AL HEALTH | CRISIS | 76 |
| 4.3 GO VIRAL! - | Study 1    | L        | 81         |           |        |    |
| 4.4 METHODS     | 82         |          |            |           |        |    |
| PROCEDURE       | 82         |          |            |           |        |    |
| Sample 83       |            |          |            |           |        |    |
| 4.5 RESULTS     | 84         |          |            |           |        |    |
| 4.6 DISCUSSION  | – Study    | 1        | 86         |           |        |    |

4.7 METHODS 86 SAMPLE 91 4.8 RESULTS 92 MANIPULATIVENESS 92 CONFIDENCE 95 SHARING MISINFORMATION 96 MOTIVATIONAL THREAT 97 **EXPLORATORY ANALYSES** 98 **ROBUSTNESS CHECKS** 99 4.9 DISCUSSION – STUDY 2 100 **4.10 GENERAL DISCUSSION** 101 BOOSTING ATTITUDE CERTAINTY THROUGH INOCULATION 102 STOPPING THE SPREAD OF MISINFORMATION 102 THE STATE OF TRUTH 102 MOTIVATIONAL THREAT AS A KEY COMPONENT OF INOCULATION 103 ROLLING OUT THE PSYCHOLOGICAL VACCINATION 104 SHORTCOMINGS AND FUTURE DIRECTIONS 105 4.11 CONCLUSION 105

### <u>5 HARNESSING POST-INOCULATION TALK TO CONFER INTRA- AND</u> INTERINDIVIDUAL RESISTANCE TO PERSUASION 107

5.1 ABSTRACT 108 **5.2 GENERAL INTRODUCTION** 109 **REVISITING COUNTERARGUING** 109 RIPPLE EFFECT OF COUNTERARGUING: POST-INOCULATION TALK 110 WHAT ARE THE UNDERLYING MECHANISMS OF PIT? 111 WHY DOES IT MATTER? 112 WHAT QUESTIONS REMAIN UNANSWERED? 112 AIMS OF THIS CHAPTER 114 **5.3 ESTABLISHING POST-INOCULATION TALK** 114 **ISSUE SELECTION 114** Methods 115 RESULTS 126 DISCUSSION - PHASE 1 131 134 CONCLUSION

5.4 SHEDDING LIGHT ON PIT'S ROLE IN ATTITUDINAL RESISTANCE 135 135 Method RESULTS 137 DISCUSSION – PHASE 2 143 **5.5 INTERINDIVIDUAL RESISTANCE THROUGH VICARIOUS INOCULATION** 147 147 METHODS 149 RESULTS DISCUSSION – PHASE 3 153 **5.6 GENERAL DISCUSSION** 156 **REVISITING AND UPDATING COUNTERARGUING** 156 IT IS ALL IN THE DETAILS – TOWARD A MORE NUANCED PORTRAIT OF PIT 158 EXTENDING THE REACH OF RESISTANCE 158

### 6 GENERAL DISCUSSION 161

| ATTITUDE CERTAINTY   | 165                |                 |  |  |  |  |
|--|--------------------|-----------------|--|--|--|--|
| UPDATING THE FUNDAMEN  | ITALS: THREAT, COU | NTERARGUING 166 |  |  |  |  |
| RESISTANCE IN THE DIGITA                                       | ALAGE 167          |                 |  |  |  |  |
| THE CASE FOR ACTIVE, GENERALISED, AND THERAPEUTIC INOCULATIONS |                    |                 |  |  |  |  |
| WHERE DO WE GO FROM H  | ere? 169           |                 |  |  |  |  |
| LESSONS LEARNED AND TA   | KE-AWAY MESSAGES   | s 170           |  |  |  |  |

### 7 REFERENCES 174

8 APPENDICES 199

APPENDIX 1 EXAMPLE TITLE 200

APPENDIX 2 201

# **TABLE OF FIGURES**

Figure 1: Landing screen Bad News (Panel A) and simulated Twitter engine (Panel B). 34 Figure 2: Median difference (post-pre) in reliability assessments of fake news items across treatment conditions with jitter (Panel A) and density plots of the data distributions (Panel B). 38 Figure 3: Median change scores (post-pre) of confidence in reliability judgments across treatment condi¬tions with jitter (Panel A) and density plots of the data distributions (Panel B). 39 Figure 4: Screenshots simulating WhatsApp conversations as fake news items (polarisation on the left, escalation on the right). 54

Figure 5: Join this Group landing page (left) and game environment (middle and right).56Figure 6: Study design flowchart. 57

Figure 7: Pre-post differences in reliability scores of fake news items between conditions. 59

Figure 8: Between conditions difference in the perceived reliability of fake items over time. 60

Figure 9: Between condition differences in updated confidence (pre-post) in reliability assessments of fake items. 61

Figure 10: Confidence scores between conditions throughout time points (pre, post, follow-up). 62

Figure 11: Difference scores in reliability, confidence, and sharing of fake news items across conditions. 63

Figure 12: Sharing of fake items over time (pre, post, follow-up) between conditions. 64

Figure 13: Path plot of mediation analysis on the relationship between condition and sharing intentions as mediated by confidence. \*p <.05. 65

Figure 14: Screenshot of in-game threat element used at the beginning of 'Go Viral!'. 79

Figure 15: Go Viral! landing page (left) and game environment (middle and right). 80

Figure 16: UNESCO infographics. 80

Figure 17: In-game survey screenshots: start of the survey (left), consent form (middle) and a social media post (right). 83

Figure 18: Bar graph of the perceived manipulativeness of fake news (left) and real news (right), averaged and per individual item. Error bars show 95% confidence intervals. *85* 

Figure 19: Examples of a manipulative (left) and real (right) social media post from the item rating task (Study 2). 88

Figure 20: Study flowchart. 90

Figure 21: Violin plot with jitter of post-pre manipulativeness scores of fake news posts (all countries combined). 93

Figure 22: Bar graphs of perceived manipulativeness of fake news and real news (UK only), by condition, for the pre-test (T1), post-test (T2) and 1-week follow-up (T3). 94

Figure 23: Path plot of mediation analysis on the relationship between condition and sharing intentions as mediated by confidence. \*p < .05. 99

Figure 24: Distribution of Go Viral! since launch in October 2020. 99

Figure 25: Study flowchart for all three phases. 119

Figure 26: Violin plot with jitter of perceived resistance scores between conditions. 127

Figure 27: Line graph of motivational and apprehensive threat (T1, T2) between conditions. 128 Figure 28: Word cloud plot for intentional pass-along messages composed at T1. 131 Figure 29: Word cloud map of frequently occurring words in actual pass-along messages (T2). 141 Figure 30: Line graph for spreading misinformation through pass-along messages between conditions (1= control; 2= inoculation). 142 Figure 31: Line graph for spreading resistance through pass-along messages between conditions. 143 Figure 32: Examples of actual pass-along messages (T2) turned into Tweet-sized treatment material (left to right; inoculation, misinformation, control condition). 149 Figure 33: Violin plot with jitter of attitude change (post-pre) between conditions. 150 Figure 34: Violin plot with jitter of perceived resistance scores (T2) between conditions. 151 Figure 35: Violin plot with jitter of self-reported counterarguing (T2) between conditions. 151 Figure 36: Violin plot with jitter of post-treatment messages that spread misinformation between conditions. 152 Figure 37: Violin plot with jitter of post-treatment messages that spread inoculation between conditions. 153

## LIST OF TABLES

Table 1: Summary of all PIT measures with scoring examples.124Table 2: Linear regressions measures of resistance, attitudes, attitude strength, and attitude certaintyas the dependent variables (T2).140

### LIST OF APPENDICES

Appendix 1: Average reliability (pre-post) judgments overall and for each fake news badge by experimental condition. 199 Appendix 2: Average confidence (pre-post) judgments overall and for each fake news badge by experimental condition. 199 Appendix 3: All 18 fake news items participants viewed pre-post by badge. 200 Appendix 4: All 18 chat screenshots (fake, real news items) participants viewed pre-post by badge. 201 Appendix 5: Appendix 1: Average reliability (pre-post) judgments for each fake news badge by experimental condition. 201 Appendix 6:Linear regression for demographics. 202 Appendix 7: Attack message on fictitious chemical presented at T1. 203 Appendix 8: Inoculation treatment used at T1. 204 204 Appendix 9: Control message on fictitious chemical used at T1. Appendix 11: Visualisation of the 'Degrees of Resistance' Index. 205 Appendix 10: Intercoder reliability analyses for content measures. 205 Appendix 12: PIT messages (T2) as the vicarious inoculation treatment. 206 Appendix 13: PIT messages (T2) that scored lowest on Resistance Index and were used instead of a traditional attack message. 208 Appendix 14: Fictitious, neutral, and unrelated treatment for the control condition. 208

# **1** INTRODUCTION

## 1.1 General introduction

# Societies' state of affairs

An engaged and informed individual is a prerequisite for any modern democracy to thrive (Cook et al., 2017b; Sandel, 1998). Social media and messaging platforms have drastically transformed how information is retrieved, shared, and assimilated across societies (Talwar et al., 2019). Some scholars

argue that the advent of the internet facilitates the democratisation of media and that it gives people the previously unprecedented power to share their views and news (e.g., Abbott, 2013; David, 2015). Additionally, with a substantial decrease in trust in media institutions and the fall of traditional editorial gatekeepers (Rhodes, 2022; Williams & Delli Carpini, 2016), more people turn to social media platforms as their news source (Gottfried & Shearer, 2016). Such sites can function as gateways through which individuals can come across and spread information without an editorial process that screens out false, fabricated, or even intentionally manipulative content (Guess et al., 2018, 2019; Lazer et al., 2018). Somewhat perplexingly and despite the advanced accessibility of information, the advent of the internet has also inadvertently made misinformation more ubiquitous (Groshek & Koc-Michalska, 2017; Wolf et al., 2021). This is in contrast with decades of science communication research which heavily relied on a model of information deficit – suggesting that it is the lack of access to facts that accounts for the prevalence of misleading information (Ecker et al., 2022; Simis et al., 2016).

To better understand the phenomenon of misinformation, it is necessary to address the absence of a clear scientific understanding of what constitutes "fake news" (and its many accompanying and often interchangeably used terms) (Vraga & Bode, 2020). Across the continuously growing misinformation literature, common approaches range from "fabricated information that mimics news media content" to content that violates the editorial norms (Pennycook & Rand, 2020). However, Traberg and van der Linden (2022) argue that the most common definitions follow an unrealistic notion of overly relying on a narrow source-based conceptualisation of "fake news" rather than acknowledging that content does not need to be entirely false to be misleading and harmful (Roozenbeek & van der Linden, 2020). The increased attention to the propagation of misinformation has not been limited to the scientific community. Indeed, in the early stages of the pandemic, the World Health Organization declared an 'infodemic' - characterised by an overabundance of false, misleading, and harmful information (Zarocostas, 2020). Similarly, in 2016, the Oxford Dictionary nominated "post-truth" as their word of the year, describing "the circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief" (Mcintryre, 2018, p.5). Research on "thought contagion" suggests that misinformation spreads rapidly across large audiences (Iyengar & Massey, 2019; Kucharski, 2016) and that false and unverified information travels faster and further than other types of information (Petersen et al., 2019; Vosoughi et al., 2018). Consequently, to understand the spread of misinformation, research has increasingly applied models from epidemiology (Cinelli et al., 2020; Kucharski, 2016; Scales et al., 2021). Indeed, the key focus of these models concerns the reproduction number at the secondary level, that is, examining the number of individuals who start posting misinformation after having come in contact with someone who was already doing so (infectious individual). Hence, it can be argued that misinformation can be approached as a viral pathogen that can infect its host and rapidly spread across individuals without direct physical contact (van der Linden, 2022).

However, its prevalence and propagation aside, does online misinformation have any real-world consequences? Researchers argue that the unprecedented scale and pace at which misinformation spreads in the digital infrastructure poses a severe threat to science, society, and the democratic process (Lewandowsky et al., 2017; van der Linden, 2022). To give a few examples, research has identified misinformation as a contributor to various societally contentious events, ranging from political elections and referenda (Bennett & Livingston, 2018; Ecker et al., 2022) and to climate change mitigation (Cook et al., 2017). Within the context of the pandemic, research has shown that the endorsement of misinformation undermines compliance with public health guidelines and decreases vaccine uptake intentions (Loomba et al., 2021; Vivion et al., 2022). Indeed, within the environment where misinformation abounds, scientific evidence is increasingly being questioned and a decrease in trust is evident (Iyengar & Massey, 2019). Furthermore, research suggests that misinformation contributes to the instigation of violence (Jolley & Paterson, 2020) and have inspired mob lynchings (Arun, 2019). In short, misinformation can have serious consequences ranging from violence and death to undermining efforts to mitigate climate change, the biggest existential threat of our time (van der Linden et al., 2017; 2021).

# The history of misinformation and its adaptations to the digital age

Of course, misinformation is not a novel concept. A quick look at the Roman emperors as well as Goebbels' efforts in spreading Nazi propaganda powerfully demonstrates the role that misinformation, whether through coin inscriptions, printed press, radio, or cinema, played in mass communication for a long time (Hekster, 2013; Herf, 2005). However, though no new concept, the spread of rumours, false information, and propaganda, reaches new levels of danger when combined with the contemporary digital information infrastructure and, more importantly, how human cognition navigates through it. For instance, the above-mentioned consequences of misinformation on climate change denial and the rejection of vaccinations suggest that misinformation is not merely a consequence of ignorance but rather, driven by psychological motives such as fear, motivated reasoning, conspiratorial thinking, and affective drivers underpinning attitude formation (Fazio et al., 2015; Hornsey et al., 2018; Nisbet et al., 2015).

More generally, research on information-seeking and processing behaviours highlights the role of psychological factors such as confirmation bias (Frenda et al., 2011; Zhou & Shen, 2021; Zhu et al., 2010), cognitive depletion (Szpitalak & Polczyk, 2014), and social cohesion (Schiefer & van der Noll, 2017). Indeed, it is argued that these processes are a function of evolutionary adaptations, allowing us to seek out people to trust and navigate an (increasingly) overstimulating world (Haselton et al., 2009; Peters, 2020a, 2020b). While these psychological mechanisms may be beneficial, scholars point toward an interplay between cognitive biases and the contemporary information infrastructure which,

in turn, reinforces and accelerates the spread of harmful information to previously unseen proportions (Murphy et al., 2020; Walter & Murphy, 2018). Building on that, research suggests that social media also highlights and amplifies moral and emotional messages, which take precedence over evidencebased decision-making (Crockett, 2017; Effron & Raj, 2020; Rathje et al., 2021). Additionally, increasingly sophisticated algorithms, 'filter bubbles' and 'echo chambers' do not only reinforce and accelerate the cognitive processes already in place, but they also organise people into digital silos of like-minded people where the absence of a 'filter-bridge' further aids the increasing polarisation of beliefs (for review, see Arguedas et al., 2022). This is particularly troubling given research that suggests that repeated exposure to misinformation makes people more likely to believe it (van der Linden et al., 2021). Therefore, what is it that makes an individual susceptible to misinformation?

Two present dominant explanations for the susceptibility to and sharing of misinformation are offered by the accounts of motivated reflection (Kahan et al., 2007) and "the classical reasoning" account of misinformation belief (Pennycook & Rand, 2020). To briefly summarise, whereas the former suggests that reasoning can increase identity-protective biases, the latter argues that a lack of "reflective openmindedness" underpins belief in misinformation and that identity-protective thinking plays a relatively minor role (Pennycook & Rand, 2020, p.197). However, neither of these theoretical accounts manage to explain susceptibility to and the sharing of misinformation in its entirety. First, the motivated reasoning account struggles to disentangle whether partisan biases are a result of motivated reasoning or selective exposure (Druckman & McGrath, 2019; Tappin et al., 2020). Similarly, a recent reanalysis of Pennycook and Rand (2019) demonstrated that while cognitive reflection was associated with enhanced truth discernment, it was not associated with partisan bias (Batailler et al., 2022). Thus, both theoretical accounts suffer from substantial shortcomings when attempting to study the exposure to, believe in, and spreading of misinformation. Indeed, researchers have called for a more "integrative account" of misinformation belief - one where, in addition to purely cognitive factors, identityprotective thinking, "myside bias", and political ideology play a central role in predicting susceptibility to misinformation (Van Bavel et al., 2021; van der Linden, 2022). In fact, when Roozenbeek and colleagues (2022) pointed out the broad variety of items, scales, question framings, and response modes when examining susceptibility to misinformation, they found that different response modes yielded similar (yet, not identical results), arguing for an "integrative" account of misinformation belief. In regard to the *sharing* of misinformation, scholars call for an urgent examination of the underlying psychological factors and warn against relying on source-based definitions of misinformation and not underestimating the damage a piece of misinformation shared by a mainstream outlet can cause (Traberg, 2022).

### Current efforts to counter misinformation

Across disciplines, the ongoing efforts to counter the spread of misinformation can be divided into four categories: legislative, technological, corrective, and educational (Haciyakupoglu et al., 2018).

To give an example, Germany introduced the Network Enforcement Act (NetzDG), obligating social media platforms to remove 'clearly illicit' content within 24 hours or face a heavy fine (Zipursky, 2019). Though this law is considered a step towards delegations of public power (Belli & Cavalli, 2019) it has also been criticised for massively damaging the basic rights to freedom of press and freedom of expression, a slippery slope that legislative efforts must tread carefully (Oliva, 2020). As a result, and in an attempt to protect their business models, many social media platforms adopted proactive actions against the prevalence of misinformation and illegal hate speech (Angelopoulos et al., 2016; Angelopoulos & Smet, 2016). Thus, these platforms adjusted their regulations, designed more sophisticated frameworks for identifying illegal content, hired additional moderators, and introduced algorithmic efforts to flag and remove harmful content. However, it is important to note that on top of being criticised for severely violating human rights (Perel & Elkin-Koren, 2015), research suggests that content moderation does not resolve the threats of misinformation but rather lead to additional 'variants' of the misinformation virus. The pivot to end-to-end encrypted private messaging apps provides one example for the adaptive nature of misinformation (Urman & Katz, 2020) and highlights the need to understand *why* misinformation arises and spreads in first place.

Furthermore, a plethora of cognitive research highlights why these technological efforts may not only be insufficient but potentially even harmful. Specifically, within attempts to correct and educate the public, two main approaches are evident - namely, reactive and proactive efforts. To begin with, reactive efforts concern the efficacy of debunking and debiasing (Lewandowsky et al., 2012). As debunking misinformation does inevitably repeat and reinforce the misinformation itself, this approach comes with several challenges. A large body of cognitive research emphasises the phenomenon of the illusory truth effect, where the mere repetition of falsehoods contributes to the perceived truthfulness of the content (Fazio et al., 2015; Pennycook & Rand, 2020). And though no consistent support for the previously feared backfire effect of corrections is evident (Ecker et al., 2019; T. Wood & Porter, 2019), debunking misinformation can be challenging in light of (politically) motivated cognition (Kahan et al., 2007; Flynn et al., 2017). Furthermore, the continued influence effect suggests that even after falsehoods have been debunked, people can continue to retrieve and rely on them from their memory (Chan et al., 2017; Lewandowsky et al., 2012). However, even when corrections are effective (MacFarlane et al., 2020), their speed and virality cannot keep up with the pace and depth at which false and unverified information can travel online (Vosoughi et al., 2018). Similarly, even if debunking and fact-checking are effective, the processing fluency, that is, the enhanced familiarity and ease with which repeated claims are perceived as true, constitutes a substantial shortfall of reactive efforts against misinformation (Wang et al., 2016). Thus, it is precisely the interplay between cognitive processes and such technological advances that can accelerate and amplify the proliferation of harmful falsehoods, even after attempts to retract, correct, and undo their harms (Murphy et al., 2020; Walter & Murphy, 2018). Coming back to the analogy of thinking of misinformation as a 'virus' and its prevalence as an 'infodemic', it becomes clear that legislations and technological tweaks (e.g., flagging, censoring, and removing content) merely deal with the symptom of this viral virus, not its cause. Consequently, leveraging psychological insights into why people fall for, believe in, and spread misinformation, to begin with, is a crucial puzzle piece to effectively combat this societal challenge. One promising alternative to reactively fighting misinformation is offered by inoculation theory.

## Origins of Inoculation Theory

When in the aftermath of the Koran War, US prisoners of war decided to remain with their captors, the assumption was that they had been brainwashed (Bernard et al., 2003). Until then, persuasion research had exclusively focused on factors that made messages more effective and regarded persuasion solely as "a facilitator for change" (Miller & Burgoon, 1972). When Lumsdane and Janis (1953) reported differing effects of message-sidedness on the effectiveness of persuasive messages, a pivotal moment was elicited in persuasion research. Contrary to limiting its focus to factors that make messages more persuasive, McGuire set out to explore "ways of producing resistance to persuasion" (McGuire, 1964, p.192). Such contemplations mark the beginning of inoculation theory, a theory often regarded as "the grandparent theory of resistance to attitude change" (Eagly & Chaiken, 1993, p.561).

Though it is true that, within the contemporary context of studying attitudinal resistance, inoculation theory can be viewed as an old one, McGuire was of course, not the first to be fascinated by resistance to persuasion. Whether it is in Aristotle's *Rhetoric* or Peacham's (Peacham, 1577) "The Garden of Eloquence", where he mentions procatalepsis (pre-emptive mention of opponents' arguments), humans have been fascinated with persuasion and the potential resistance to it for a long time. Indeed, legal scholarship points toward the phenomenon of "stealing thunder" (McElhaney, 1987) and emphasises that "If you don't [divulge the information], your opponent will, with twice the impact." (Mauet, 1992, pp.47-48). In fact, McGuire notes in the presentation of his first set of studies on what would later be regarded as the beginning of inoculation theory that "there are many people investigating resistance to persuasion, only they [...] haven't always been aware of it" (McGuire, 1964, p.192). However, by basing it on the biological analogy of an immunisation process, inoculation theory is arguably the first to empirically study *how* and *which* psychological mechanisms underpin resistance to persuasion (Compton, 2013).

### Theoretical foundations

Inoculation theory (McGuire, 1964) posits that similarly to how injecting a weakened dose of pathogen leads to the production of antibodies, exposure to weakened persuasive arguments activates "attitude-bolstering" protective responses against future persuasion attempts. Consistent with the biological analogy, McGuire identified inoculation-induced resistance to persuasion by establishing the two theoretical pillars – threat and counterarguing.

These two mechanisms describe the forewarning or threat of an imminent counter-attitudinal attack and the pre-emptive refutation to provide arguments with which individuals may protect their beliefs in the future. Traditionally, inoculation treatments would therefore elicit implicit threat (later, explicitly by including forewarnings) and were followed by a two-sided refutational message which provides sufficient pre-emptive refutations to model the counter-arguing process and motivate the generation of "attitude-bolstering" arguments against future persuasion attempts (Banas & Rains; Compton, 2013). Hence, by incorporating an affective (threat) and cognitive (counterarguing) component, the inoculation treatment uses challenges that are strong enough to trigger the mind's defence system to build an arsenal of belief-protecting 'mental antibodies' but not so persuasive that they overwhelm it (Compton et al., 2016; Compton & Pfau, 2009; McGuire, 1964). Some scholars have speculated whether affective threat alone suffices to confer resistance (Freedman & Sears, 1965; Wyer, 1976). Research has shown that forewarning accompanied by "refutational pre-emption" confers resistance more effectively than forewarning alone (McGuire & Papageorgis, 1964; van der Linden et al., 2017). In sum, while threat functions similar to the injection of a weakened virus, counterarguing as a process is believed to mimic antibodies, attacking and weakening the antigens (Compton, 2013). It becomes clear that the basic model of inoculation theory is tightly connected to medical inoculations. And while threat and counterarguing fit the analogic nature of the theory (Holyoak & Thagard, 1995), early research lacked empirical support and instead, their key components were merely assumed for the most part (Compton & Pfau, 2005).

Research eventually provided clarifications for the role of threat and counterarguing (Pfau & Burgoon, 1988), though the fact that McGuire's early work did not advance beyond cultural truisms, that is, "beliefs that are so widely shared within the person's social milieu that [the person] would not have heard them attacked, and indeed, would doubt that an attack were possible" (McGuire, 1964, p. 201), impeded its applicability. This was mostly driven by the "germ-free ideological environment" (McGuire, 1964, p.200) that such truisms offered, which allowed researchers to test the efficacy of inoculation treatments on non-political issues that people had likely never been attacked before (e.g., benefits of teeth brushing and penicillin).

Similarly, much of inoculation research limited its application to individuals that held attitudes congruent with the target topic of the study. In short, McGuire's commitment to the analogical foundations of inoculation theory kept most of its application contextually bound. As the medical analogy of inoculation research posits a preventative strategy, it is unsurprising that most of the research has constrained itself to *prophylactic* inoculation treatments, neglecting those "already afflicted" (Wood, 2007). In fact, when Wood attempted to study the effects of inoculation messages on differing pre-existing attitudes, it was hypothesised that the treatment would be ineffective or potentially backfire. Surprisingly, the findings suggested that compared to the control group, individuals with differing attitudinal predispositions were moved in the advocated direction of the inoculate individuals with opposed, neutral, and supporting attitudes alike. Since then, inoculation

theory has gained much renewed scholarly interest and the efficacy of inoculation treatments has been demonstrated in a plethora of topics ranging from peer pressure on alcohol use (Godbold & Pfau, 2016) and animal testing (Nabi, 2003b) to support for U.S. involvement in the Iraq War (Pfau et al., 2008). In short, a large body of research gives reason to surmise that inoculation treatments are not contextually bound and are "appropriate for any context where strongly held attitudes are vulnerable to challenge" (Pfau et al., 2001, p.252). Since then, research has pointed towards numerous mediators, moderators, varying outcomes, and further deviations from the analogy (for systematic review, see Banas & Rains, 2010; Compton et al., 2021), highlighting inoculation theories' complicated relationship with its analogical namesake. And as noted by Compton and Pfau (2005), although inoculation theory is a mature theory of persuasion, it is "far from retiring" (p.136). Instead, Compton (2013) emphasises that the early theorising should not function as a "theoretical point of departure" (McGuire, 1964, p.222). Thus, the present doctoral thesis aims to recognise, challenge, and ultimately advance the inherent assumptions and applications of inoculation theory.

### Theoretical advancements and new avenues

The last two decades led to inoculation theory undergoing extraordinary growth. Moving away from the monolithic and 'germ-free' setting of traditional inoculation studies and reviewing its efficacy on contested issues has emphasised the contextual boundlessness of inoculation treatments. Indeed, a meta-analysis conducted by Banas and Rains (2010) has both reinforced and challenged the traditional thinking about resistance through inoculation. A significant shift in the theoretical foundations of inoculation beyond the boundaries of the analogy and thereby, moved beyond a process that is inherently and exclusively pre-emptive (Compton, 2019). Indeed, even in the earlier stages of the theory, scholars have questioned the analogy's hold on inoculation theory. Pryor and Steinfatt (1978) pointed out fundamental discrepancies between the theory and its analogy. They argued that some dimensions of attitudinal inoculation, such as explicit forewarning, did not align with the medical analogy and that the analogy disproportionately emphasised the cognitive processes believed to be underpinning the inoculation process.

However, the authors called for inoculation scholarship to rethink the logic derived from the analogy rather than the analogy itself (Compton, 2019). Accordingly, research started moving away from cultural truisms and increasingly demonstrated the efficacy of inoculation treatments in conferring resistance against contested issues and across individuals with differing pre-existing attitudes (Banas & Rains, 2010; Wood, 2007). In many ways, the context of online misinformation is providing the perfect

setting to test, challenge, and extend the theoretical boundaries and applications of inoculation theory. Consistent with the biological foundations of the theory itself, scholars have compared misinformation to a virus (van der Linden, 2022). Additional to reviewing and predicting the spread of 'the virus' through an epidemiological lens (Cinelli et al., 2020; Kucharski, 2016; Scales et al., 2021), research examined whether *proactively* conferring resistance against misinformation was possible (van der Linden et al., 2017). In their seminal study, van der Linden and colleagues tested the efficacy of a traditional text-based inoculation treatment against the Global Warming Petition Project, one of the most potent online misinformation campaigns about climate change (arguing that over 31,000 scientists have signed a petition that there is no evidence for global warming).

To do so, participants (N = 2167) were randomly assigned to one of five conditions where they were either exposed to just an infographic emphasising the scientific consensus (97%) on human-caused global warming, a 'false-balance' condition (scientific consensus + misinformation), and one brief (forewarning about politically motivated groups followed by scientific consensus) and one detailed inoculation message (in-depth pre-emptive refutation of the petition, e.g., fake and uncredible signatories). The results found initial support for the effectiveness of both the brief (d=0.33) and detailed (d=0.75) inoculation messages in conferring resistance against online misinformation and even bolstering beliefs about the scientific consensus on climate change. Furthermore, these results have since been confirmed by two pre-registered replication studies and were shown to persist even when exposure to the misinformation message was delayed by one week (Maertens, Anseel, & van der Linden, 2020; Williams & Bond, 2020). Van der Linden and colleagues' (2017) study spearheaded the application of inoculation to numerous cases of misinformation about highly contested real-world issues ranging from vaccine hesitancy to political radicalisation (Compton et al., 2021; Lewandowsky & Yesilada, 2021; Steenbuch Traberg, 2022). Since then, inoculation scholarship begun seeing several key theoretical innovations (Compton et al., 2020; van der Linden et al, 2021).

Indeed, researchers have called for a distinction between *prophylactic* and *therapeutic* inoculation approaches (Compton, 2019; Compton et al., 2021). That is, inoculation can be fully pre-emptive (prophylactic) when (i) people hold attitudes

congruent with the treatment message and (ii) have not yet been exposed to persuasive or manipulative messages on the issue topic. On the contrary, therapeutic inoculation treatments – similarly to therapeutic medical vaccines – describe the administration to those 'already afflicted' (Compton, 2019; Wood et al., 2007). Consistent with the analogical reasoning, Compton and colleagues (2020) note that this distinction is not necessarily a departure from the analogy but, instead, consistent with the 'incubation period' where medical vaccines would still be effective and argue that inoculation treatments could confer resistance to attitudes that have had prior exposure to manipulation (but were not completely manipulated). Regardless, while these distinctions may not matter for the practical use of inoculation interventions, they are needed for the purpose of theory development. Especially with gaps in the understanding of the core constructs and mechanisms underpinning the inoculation process remaining, theoretical clarity is critical (Compton, 2013; Compton, 2019).

In a similar vein to inoculation research moving beyond cultural truisms, the pivot from prophylactic to therapeutic inoculation interventions opens new dimensions in which resistance to persuasion may be tested, conferred, and spread. Yet, both inoculation research, as well as persuasion research in general, has so far largely assumed that successful resistance to persuasion is reflected by the valence and extremity of attitudes remaining unchanged (McGuire, 1964; Z. Tormala & Petty, 2002; Zuwerink & Devine, 1996) – a notion that was somewhat passively adopted by more recent research on therapeutic inoculation interventions. Consequently, well-established attitudinal ascendents, such as attitude certainty, remain largely neglected within the context of both prophylactic and therapeutic inoculation treatments. Particularly in regard to detecting and resisting misinformation, attitude certainty is arguably crucial for three reasons.

First, confidence judgements determine whether individuals act on their initial (truth) judgements of information or whether they engage in additional informationseeking behaviours (Berner & Graber, 2008; Meyer et al., 2013). Secondly, research finds that the level with which one confidently holds their attitude affects their willingness and ability to defend and advocate for their beliefs, even if the issue itself is a contested one (Lin & Pfau, 2007; Tormala & Petty, 2004). Hence, it could be argued that individuals who are accurately confident in their ability to assess the veracity of online content will less likely fall for and share misinformation (Basol et al., 2020). And lastly, especially in the absence of other source cues, confidently expressed opinions are perceived as more trustworthy and competent (e.g., Tenney et al., 2008), thus, further highlighting the potential role attitude certainty may play in the generation, strengthening, and spreading of resistance.

Thirdly, traditional inoculation treatments have predominantly employed issue-same messaging strategies that pre-emptively debunked ('pre-bunked') the same arguments within the same issue topic as the subsequent attack message (Allen, 2009; McCroskey et al., 1972). Occasionally, the treatment conditions were tested for their effectiveness in conferring a "blanket of protection", that is, conferring resistance against different arguments within the same issue topic (Parker et al., 2016) or "cross-protection" against untreated yet related topics (Parker et al., 2012a). However, it could be argued that this issue-based approach significantly limits both the scalability of inoculation interventions and a more nuanced understanding of the analogy's boundary conditions (Bonetto et al., 2018; Roozenbeek & van der Linden, 2019b). Particularly within the context of online misinformation, two shortfalls of such an approach become apparent: persuasive arguments within the same issue topic are constantly changing in form, modality, and content (Adriani, 2019) and individuals would have to be inoculated against every argument within every topic to be adequately equipped against online misinformation. Much like generalised vaccines (e.g., MMR vaccine) that can successfully immunise against a set of viruses, recent research has demonstrated the efficacy of generalised inoculation treatments (Roozenbeek & van der Linden, 2019; Roozenbeek & van der Linden, 2020; Basol et al., 2021) by developing inoculation interventions which confer resistance against the common manipulation techniques that underpin misinformation itself.

Specifically, these techniques are partially derived from NATO's report "Digital Hydra", which outlines various forms of misinformation strategies as well as growing research on the deceptive strategies (Bertolini & Aiello, 2018; Brady, Wills, Jost, Tucker, & van Bavel, 2017; Cook et al., 2017b; Goga, Loiseau, et al., 2015; Goga, Venkatadri, et al., 2015). The process of inoculating against underlying strategies used by a whole range of misinformation is an example of conferring a "blanket of protection" against the misinformation 'virus'. That is, inoculating individuals against one strain offers immunisation against related yet different strains of the same misinformation techniques. Moreover, the gradual and weakened "dose" of misinformation strategies paired with the task of actively generating counterarguments demonstrates a critical step toward

generalised, therapeutic, and scalable inoculation interventions (Basol et al., 2021; van der Linden, 2022). Yet, although this research provides evidence for the effectiveness of therapeutic and generalised inoculation treatments against misinformation, these underlying mechanisms facilitating such effects, our current scientific understanding, as well as its potential applicability, remains vastly underexplored.

Forth, inoculation research has predominantly proposed a two-sided approach to counterarguing (Compton, 2013). Namely, the assumption was that exposure to refutational counterarguments in the treatment message extended to counterarguing as a process. That is, inoculated individuals begin to raise and refute arguments on their own after treatment exposure (Compton & Pfau, 2005; McGuire, 1964). This notion of having counterarguing modelled and subsequently, continued after treatment exposure, proposed a dynamic interplay which, though consistent with the analogical foundations of inoculation, remains mostly assumed (Banas & Rains, 2010; Compton, 2013). When counterarguing was explored as a process by requiring participants to make a list of arguments in support of their position on the issue topic, no effect of inoculation on counterarguing was found (Papageorgis & McGuire, 1961). However, by revisiting the original prediction that actively letting participants generate their own "mental antibodies", recent research has tested interactive inoculation treatments where participants are prompted to proactively counterargue against manipulation strategies (McGuire & Papageorgis, 1961; Roozenbeek & van der Linden, 2019). In doing so, research began exploring the benefits of 'active' versus 'passive' inoculation treatments (Roozenbeek & van der Linden, 2018). The key distinction here is that contrary to the traditional "passive" provision of counter-arguments which individuals can adopt and use when encountering (manipulative) persuasion at a later stage, individuals are *actively* engaging in the process of generating counterarguments themselves. Although these are substantial steps toward thinking about and implementing inoculation treatments, little is currently known about the differences between active and passive inoculation treatments (Compton et al., 2021). Similarly, while these treatments operationalize counterarguing differently both lack a clear understanding of the mechanisms underpinning it. Consequently, taking a closer and more critical look at counterarguing, one of the assumed theoretical pillars of inoculation, is crucial to counteract the current theoretical opaqueness and identify how resistance through inoculation manifests and may be spread.

Lastly, until recently, inoculation scholarship considered one of its core theoretical concepts, counterarguing, as a distinctly subvocal process. That is, counterarguing was believed to be the inoculated individuals' intrapersonal dialogue and grappling with the arguments raised in the attack message. By treating it as an exclusively internal process, research on counterarguing has predominantly limited itself to individual differences that mediate such intrapersonal communication (Compton & Pfau, 2009). However, Ivanov and colleagues (2012) proposed that fully vocalised counterarguing through actual talk might play an equally important role in resistance. By suggesting that counterarguing may be simultaneously subvocal and vocal, the authors took an important step toward thinking of vocalised counterarguing, or post-inoculation talk (PIT), as an intrapersonal and interpersonal process (Compton & Pfau, 2009). Indeed, limited work on post-inoculation talk has emerged since then (Dillingham & Ivanov, 2016a; Ivanov, Parker, et al., 2018; Ivanov, Sellnow, et al., 2018). While some work on post-inoculation talk (PIT) has been conducted, substantial theoretical and practical gaps remain that prevent the use of PIT to its full potential.

For example, the few existing studies on PIT instructed participants to engage in or withhold from post-inoculation talk. Doing so has prevented any clear conclusions regarding whether post-inoculation talk occurs organically and voluntarily after treatment exposure. Similarly, the talk was predominantly assessed by self-reported and recalled frequency of conversations and the number of conversational partners. In short, until now, inoculation scholarship approached vocalised counterarguing in form of post-inoculation talk primarily by instructing individuals to talk and subsequently comparing whether any quantitative differences occurred between treatment conditions. Research must move beyond the current snapshot approach of post-inoculation talk and establish whether and how intensely it occurs organically, whether engaging in PIT impacts their beliefs and attitudes, and, perhaps more importantly, what inoculated individuals talk *about*.

### Pushing the boundaries of Inoculation Theory

Although Ivanov and colleagues (2016) aimed to examine the content of post-inoculation talk, their focus was limited to the effects of PIT on the *spreader*. Contrastingly, particularly within the context of misinformation, various scholars have highlighted the necessity to explore ways to spread attitudinal resistance from one person to another (Basol et al., 2021; Compton et al., 2021; Lewandowsky & van der Linden, 2021a). This doctoral thesis, at least at the time of the write-up, is the first to propose assessing the efficacy of post-inoculation talk by *vicariously inoculating* recipients of post-inoculation talk and therefore, taking a novel step towards spreading resistance *between* 

individuals. Furthermore, research has yet to identify the prerequisites for verbally passed-on inoculation messages to be effective. That is, whether PIT can function as an inoculation treatment and, if so, whether it needs to mimic core components of traditional inoculation treatments. In other words, to what extend does post-inoculation talk need to meet the prerequisites of inoculation treatments (i.e., threat and counterarguing) to effectively pass-on resistance from one individual to another? Only once research establishes a more nuanced understanding of whether inoculated individuals organically engage in post-inoculation talk, what they talk about, how often they do so, and what role PIT plays in generating (intraindividual) as well as passing on (interindividual) resistance, can research begin exploring the possibility of post-inoculation talk as a promising pathway towards psychological herd immunity against persuasion (Compton et al., 2021; Lewandowsky & van der Linden, 2021a; van der Linden, 2022). This is of particular importance given that once enough individuals are 'psychologically vaccinated', the spread of misinformation will be curbed and will not spread within a population. In short, instead of attempting to catch up with, correct, and undo the harms of misinformation, this doctoral thesis argues that post-inoculation talk could play a crucial role in spreading resistance against misinformation from one individual to another. If being vicariously inoculated proves to be an effective mean to confer inter-individual resistance against misinformation, inoculation may have a chance at keeping up with, if not outpacing, the speed and depth at which harmful falsehoods travel.

To summarise, despite the prodigiously growing literature, fundamental aspects of the mechanisms that facilitate attitudinal resistance remain unanswered. It can be argued that this neglect stems from an overly cognitive approach to the conceptualisation and implementation of inoculation theory, neglecting the role of affect and actual behaviour (Pfau, 2001). Instead, as posited by the literature on attitudes, a tripartite theoretical approach that includes the components of affect, cognition, and behaviour should be pursued (Eagly & Chaiken, 1992). Inoculation research needs to move beyond the mere demonstration of its efficacy and establish a scientifically rigorous understanding of the mechanisms that explain why and how inoculation treatments are effective in conferring resistance. To do so, the above-mentioned gaps in the literature will need to be addressed. Irrespective of differing views on whether and how anchored inoculation theory should remain to its biological namesake, inoculation research is at a pivotal moment (Compton, 2019; Compton et al., 2021; Wood, 2007). Further examining and advancing the above-mentioned new avenues could result in significant theoretical and practical innovations. Critically reviewing and rethinking core aspects of how inoculation research defines and operationalises threat, counterarguing, and resistance itself will allow for inoculation theory to be well-positioned within the context of the post-truth era (Compton, van der Linden, Cook, & Basol, 2021). Only once a clearer understanding of the pre-requisites, driving factors, and boundary conditions of active, generalisable, and therapeutic inoculation treatments exists can we start taking decisive and effective steps towards psychological herd immunity against misinformation. This doctoral thesis hopes to contribute to such aspirations by advancing the current scientific understanding of how attitudinal resistance is build, strengthened, and shared.

### Outline of Doctoral Thesis

Across the next 5 chapters, this doctoral thesis will offer a multi-layered approach to thinking about, testing, and enhancing resistance to persuasion. Thus, I will review the possibilities of conferring resistance through a variety of treatment forms (e.g., active vs. passive), across various contexts (e.g., generalised vs. specific), on different platforms (e.g., simulating Twitter, WhatsApp, and intra- and interpersonal talk) with the aims to expose the mechanisms that allow inoculation to confer resistance and establish how such resistance may be build, strengthened, and shared. Accordingly, Chapter 2 will begin by examining the effectiveness of active, generalised, and therapeutic inoculation treatments and will review the role of attitude certainty in the inoculation process. Subsequently, Chapter 3 will further explore the role of attitude certainty in the strengthening of resistance and build on these findings and extend the exploration of gamified inoculation treatments to the context of end-to-end encrypted messaging applications and the distinct psychological factors that accompany them (e.g., group dynamics, rumour, and gossip). Next, Chapter 4 will address the shortcomings of the previous chapters and directly compare the effectiveness of active and passive forms of inoculation interventions against COVID-19-specific misinformation. This chapter will also critically revisit the role of threat, one of two key theoretical components of inoculation theory, and identify ways to enhance our current scientific understanding and operationalisation of the role of threat in the resistance process. In doing so, these chapters will highlight the processes that allow inoculationinduced resistance to be generated, strengthened, and spread. Lastly, Chapter 5 will examine threat's counterpart - counterarguing. Here, a three-phased experiment will explore whether vocalised counterarguing, in form of post-inoculation talk (PIT), offers a feasible way toward psychological herd immunity. To do so, this chapter will first establish whether inoculated individuals organically and voluntarily engage in post-inoculation talk and if so – what do the quantity, depth, and content of their conversations look like? Additionally, this chapter will assess whether engaging in PIT, in turn, has any beneficial effects on the inoculated individual (e.g., strengthening attitudes). Lastly, this chapter will examine whether post-inoculation messages composed by inoculated individuals can function as a substitute for traditional inoculation treatments and vicariously inoculate the recipients of PIT. Crucially, by pitting inoculation against misinformation and assessing whether vicariously inoculated individuals pass on inoculation-congruent content instead of misinformation, this chapter will explore a potential pathway towards societal immunity against misinformation through inoculation.

Jointly, these studies aim to review, challenge, and advance our current understanding of inoculation theory, its core theoretical constructs, its operationalisation, and its application. Specifically, the research presented in this thesis contributes to the current (lack of) understanding of the role of attitude certainty, motivational threat, and post-inoculation talk in the inoculation process and even goes as far as questioning as to what constitutes as realistic resistance in the digital age. By challenging the foundations and its traditional interpretations alike, I hope to shed some light on how far inoculation scholarship has come and where it can go from here. Therefore, across the next five chapters, this doctoral thesis aims to lay out a pathway for resistance to persuasion to be build, strengthened, and spread. Lastly, the highly applied nature of this thesis provides unique insights, for scholars and policy-makers alike, into the efficacy, shortfalls, and future directions of inoculating intervention against pressing societal challenges.

# 2 THE ROLE OF ATTITUDE CERTAINTY IN

# **RESISTANCE TO PERSUASION**

**Published as:** Basol, M., Roozeneek, 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. DOI: <u>http://doi.org/10.5334/joc.91</u>

## 2.1 Abstract

Recent research has explored the possibility of building attitudinal resistance against online misinformation through psychological inoculation. The inoculation metaphor relies on a medical

analogy: by pre-emptively exposing people to weakened doses of misinformation cognitive immunity can be conferred. A recent example is the Bad News game, an online fake news game in which players learn about six common misinformation techniques. We present a replication and extension into the effectiveness of Bad News as an anti-misinformation intervention. We address three shortcomings identified in the original study: the lack of a control group, the relatively low number of test items, and the absence of attitudinal certainty measurements. Using a 2 (treatment vs. control)  $\times$  2 (pre vs. post) mixed design (N = 196) we measure participants' ability to spot misinformation techniques in 18 fake headlines before and after playing Bad News. We find that playing Bad News significantly improves people's ability to spot misinformation techniques compared to a gamified control group, and crucially, also increases people's level of confidence in their own judgments. Importantly, this confidence boost only occurred for those who updated their reliability assessments in the correct direction. This study offers further evidence for the effectiveness of psychological inoculation against not only specific instances of fake news, but the very strategies used in its production. Implications are discussed for inoculation theory and cognitive science research on fake news.

### 2.2 Introduction

### 2.2.1 Persuasion Research on Attitude Certainty

Much of early persuasion research has given priority to studying incidences where persuasion is successful, prioritising an understanding of what makes messages *more persuasive*. Around the early 60s, an increasing number of scholars dedicated themselves to understanding the process of defending one's attitudes against persuasive messages – that is, attitudinal resistance to persuasion. Since then, research has demonstrated that people tend to resist persuasive attempts when they are forewarned about someone's manipulative intent (Papageorgis, 1968; Petty & Cacioppo, 1979), when their

perceived freedom is threatened (Brehm, 1966), and when attitudes are held strongly (Petty & Krosnick, 1995). Indeed, general persuasion research has identified a number of distinct factors and mechanisms that underpin resistance to persuasion (Petty, Tormala, & Tucker, 2004; Banas & Rains, 2010). To exemplify, studies suggest that bolstering initial attitudes, derogating the credibility of a persuasive message, or experiencing negative affect (e.g., anger) play important roles in resisting persuasion (Lewan & Stotland, 1961; Tannenbaum et al., 1996; Ahluwalia, 2000).

Another powerful ingredient to resistance to persuasion is attitude certainty, a dimension of attitude strength, which refers to the "degree to which an individual is confident that his or her attitude toward an object is correct" (Krosnick et al., 1993, p.1132). A large body of research has established antecedents that offer glimpses into the unique ways in which individual and contextual factors affect attitude certainty. To give an example, Petrocelli and colleagues (2007) argued that two distinct components, which are referred to as attitude clarity and attitude correctness, warrant a more nuanced conceptualisation of attitude certainty. Whereas attitude clarity refers to how confidently one is aware of their attitude towards an issue, attitude correctness relates to the subjective correctness and validity with which an attitude is held. Indeed, Petrocelli and colleagues (2007) suggest that despite their highly correlated nature, these two components should be examined independently of one another. Other research highlighted the relationship between attitude accessibility and attitude certainty, such that repeatedly expressed attitudes were held with higher levels of certainty than attitudes that were not (Holland et al., 2003).

However, current conceptualisations of attitude certainty assume that the consequences of attitude certainty occur regardless of how certainty is reached or established. And though research suggests that the sense of conviction with which an attitude is held predicts behavioural intentions, strengthens resistance to persuasion, and persists over time (e.g., Fazio & Zanna, 1978; Bassili, 1996), inoculation theory, arguably the most prominent account of empirically testing attitude resistance, has neglected attitude certainty for the most part (Pfau et al., 2005; Tormala & Petty, 2004). In short, inoculation theory (McGuire, 1964), suggests that analogous to the process of a medical vaccination, the two key components threat and counterarguing, model and continue the generation of attitude-bolstering "mental antibodies" against future persuasive attempts. The logic behind it suggests that initial exposure to mild forms of persuasion triggers a heightened perception of one's attitudinal vulnerability, which, in turn, motivated the individual to take on pre-emptive refutations offered in the treatment messages and build mental defences against future attacks on the issue topic.

### 2.2.2 Inoculation Research on Attitude Certainty

Inoculation theory implies that the initial attack and subsequent heightened sense of attitudinal vulnerability led attitude certainty to decrease. However, after attitude certainty was initially shaken by the threat component of the treatment message, the generation of and practice with counterarguments was assumed to increase one's confidence to resist future forms of persuasion

(McGuire, 1964). However, this notion points to attitude certainty within the context of resisting, rather than how it may affect the certainty with which an attitude is held per se (Tormala & Petty, 2004). Interestingly, in both research on inoculation theory as well as general research on resistance to persuasion, the predominant assumption appears to be that successfully resisting equals attitudes remaining entirely unchanged. On the other hand, more recent research found that resisting persuasive attacks had a strengthening effect on the target attitude (Tormala & Petty, 2002). Specifically, the authors demonstrated that, under some circumstances, successfully resisting persuasive attempts has a boosting effect on the confidence with which the target attitude is held. Additionally, the authors proposed a novel and metacognitive account of how individuals' awareness of their resistance affected their attitude certainty. Importantly, Tormala and Petty found that the more confident individuals became in their attitudes, the more these attitudes predicted behavioural intentions. Since then, inoculation research identified various unintended benefits (or "by-products") of treatment messages, ranging from increasing perceived self-efficacy, attitude accessibility, and attitude certainty (Compton & Pfau, 2005; Ivanov et al., 2009). Indeed, research suggests that on top of attitude certainty's effect on resistance, it also strengthens attitude persistence and predicts behavioural intentions (Brügger & Höchli, 2019; Tormala & Petty, 2004). Thus, these findings emphasise the potential role of attitude certainty in strengthening beliefs and predicting behavioural intentions, which, in turn, are most predictive of actual behavioural actions (Maki et al., 2019).

# Pushing the theoretical boundaries: Inoculating against misinformation

However, it is essential to note that much of the early inoculation work was constrained to "cultural truisms", that is, beliefs that are so widely held within the social milieu, that the very notion of challenging them seems implausible and unlikely (McGuire, 1964). In the real world, however, people will often hold very different, at times even contradictory pre-existing beliefs about a particular issue. As a consequence, the current understanding of the role that attitude certainty plays on the inoculation process with contested issues, or simply topics that are either still evolving or that people have differing opinions on, is somewhat fragmented still. More recently, and because the spread of harmful content in online networks bears a close resemblance to the manner in which a virus replicates (Kucharski, 2016), inoculation theory has been applied to the context of online misinformation. In fact, research emphasises the efficacy of inoculation treatments in the context of disinformation campaigns about climate change (Cook et al., 2017; van der Linden, Leiserowitz, et al., 2017), political radicalisation (Lewandowsky & Yesilada, 2021), and conspiracies about the COVID-19 pandemic (Basol et al., 2021). Since people have likely come across misinformation about most real-world settings prior to treatment exposure, from a theoretical point of view, we cannot speak of purely prophylactic inoculation. Instead, just as medicine has advanced to distinguish between *prophylactic* and therapeutic vaccines, therapeutic inoculation approaches can still confer protective benefits even

among those already "afflicted" by boosting immune responses in the desired direction (Compton, 2019).

### Finding an Antidote: Gamified Inoculation

More recently, Roozenbeek and van der Linden (2019) have pointed toward novel ways to apply gamified inoculation interventions against online misinformation. Here, participants enter a simulated social media environment (Twitter) where they are gradually exposed to weakened "doses" of misinformation strategies and actively encouraged to generate their own content. The intervention is a free social impact game called Bad News (www.getbadnews.com; Figure 1A), developed in collaboration with the Dutch media platform DROG (DROG, 2018), in which players learn about six common misinformation techniques (impersonating people online, using emotional language, group polarisation, spreading conspiracy theories, discrediting opponents, and trolling, Figure 1B). These strategies are partially derived from NATO's report "Digital Hydra", which outlines the various forms of misinformation strategies as well as growing research on the deception strategies (Bertolini & Aiello, 2018; Brady et al., 2017; Cook et al., 2017; Goga, Loiseau, et al., 2015; Goga, Venkatadri, et al., 2015).





Figure 1: Landing screen Bad News (Panel A) and simulated Twitter engine (Panel B).

The purpose of the game is to produce and disseminate disinformation in a controlled environment whilst gaining an online following and maintaining credibility. Players start out as anonymous netizens and eventually rise to manage their own fake news empire. The theoretical motivation for the inclusion of these six strategies is explained in detail in Roozenbeek and van der Linden (2019) and covers many common disinformation scenarios including false amplification and echo chambers. Moreover, although the game scenarios themselves are fictional they are modelled after real-world events. In short, the gamified inoculation treatment incorporates an active and experiential component to resistance-building.

### Shortcomings and Remaining Gaps

Although the study provided preliminary evidence that the game increases people's ability to detect and resist a whole range of misinformation (in the form of deceptive Twitter posts), the study suffered from a number of important theoretical and methodological limitations. For example, although the original study (Roozenbeek & van der Linden, 2019) did include various "real news" control items, it relied on a self-selected online sample of approximately 15,000 participants in a pre-post (within) gameplay design, and therefore, lacked a proper randomized control condition. This is important because there could be a secular trend so that people downgrade their reliability ratings of the fake tweets (pre-post) regardless of what intervention they are assigned to. Second, because the testing happened within the game environment, the original study only included a limited number of fake news items (one survey item per misinformation technique). Third, on a theoretical level, the study only looked at reliability judgments and thus could not determine how confident or certain people actually were in their beliefs. While there is some research emphasising the role of confidence in identifying misinformation and belief in conspiracy theories (Halpern et al., 2019; Hinsley et al., 2022; Ognyanova et al., 2020), little is known about how one's attitude confidence affects resistance against misinformation.

How then, can this somewhat fragmented understanding of attitude certainty's role in the inoculation process be applied to the current societal threat of online misinformation? Do inoculation treatments leave individuals less or more confident in their ability to spot and resist misinformation? Does attitude confidence have any impact on their actual ability to resist misinformation? Additionally, it is unclear whether the same theoretical mechanisms that facilitate prophylactic inoculation (e.g., confidence in defending one's beliefs) also boost the efficacy of therapeutic inoculation. Addressing these theoretical and practical questions will be crucial to maximising the potential efficacy of generalised and therapeutic inoculation interventions against misinformation. To summarise, this chapter addresses three key shortcomings in the original research by 1) including a randomized control group, 2) adding a larger battery of items, and 3) identifying *whether* inoculation-induced attitudinal resistance increases attitude certainty, and, in turn, how this might affect *how* attitude resistance through inoculation is build and strengthened.

### 2.3 Methods

### Participants and Procedure

This study employed a 2 (*Bad News*. vs. Control) \* 2 (pre-post) mixed design to test the efficacy of active (gamified) inoculation in conferring attitudinal resistance to misinformation. The independent variable consisted of either the treatment condition in which participants played the *Bad News* game or a control condition in which participants were assigned to play *Tetris* (to control for gamification;

*Tetris* specifically was chosen because it is in the public domain and requires little prior explanation before playing).

Following Roozenbeek and van der Linden (2019), the dependent variable consisted of an assessment of the reliability of 18 misinformation headlines in the form of Twitter posts (see appendix). As the *Bad News* game covers six misinformation techniques, three items per technique were included<sup>1</sup>. These Twitter posts were created to be realistic, but not real, both to avoid memory confounds (participants may have seen "real" fake news headlines before) and to be able to experimentally isolate the misinformation techniques. Taking into account the average inoculation effect reported in previous research (Roozenbeek & van der Linden, 2019), a priori power analysis was conducted with G\* power using  $\alpha = 0.05$ , f = 0.26 (d = 0.52) and power of 0.90 with two experimental conditions. The minimal sample size required for detecting the main effect was approximately 158. A total of 197 participants were recruited through the online crowdsourcing platform, *Prolific Academic*, which has been reported to produce higher data quality than *MTurk* (Peer et al., 2017). Consenting participants (58% male, modal age bracket = 18–24, 20% higher educated, 61% liberal, 80% white<sup>2</sup>) completed the survey, were debriefed, and paid £2.08 in compensation. This study was approved by the Cambridge Psychology Research Ethics Committee (PRE.2018.007).

A plug-in was created so that the game could be embedded in *Qualtrics*, and pre-post testing could take place outside of the game environment to further enhance ecological validity. Upon giving informed consent, participants were randomly presented with 18 fictitious Twitter posts and on a standard 7-point scale, reported how reliable they received each post to be and how confident they were in their judgements. Subsequently, participants were randomly assigned to a condition. In the inoculation condition participants (n = 96) were asked to play the "*Bad News*" game for about 15 minutes. Participants were assigned a password for completion which they could only receive after completing the final level (badge). Participants (n = 102) in the control condition played *Tetris* for 15 minutes in the same manner. After treatment exposure, all participants were asked to complete the same set of outcome measures.

<sup>&</sup>lt;sup>1</sup> In the original study by Roozenbeek and van der Linden (2019), only six items were included. We included the original items plus two new ones for each badge using the same approach.

<sup>&</sup>lt;sup>2</sup> Socio-demographics (except for ideology) were answered by 52% (n = 104) of the 197 participants.
### Measures

### Perceived Reliability

To assess participants' perceived reliability, a single-item measure was presented alongside 18 (6\*3) fake Twitter posts (example item polarization; "*New study shows that right-wing people lie more often than left-wing people*"). Participants reported the perceived reliability of each post on a 7-point Likert-scale from not reliable at all (1), neutral (4) to very reliable (7). Following Roozenbeek and van der Linden (2019), to form a general fake news scale of perceived reliability, all 18 fake news items were averaged. An initial reliability analysis suggested good internal consistency (M = 3.17, SD = 0.85,  $\alpha = 0.84$ ) of the 18-item fake news scale. A subsequent exploratory principal component analysis (PCA) was also run on the fake news items. According to the Kaiser criterion, results indicated that the items clearly loaded on a single dimension with an eigenvalue of 3.15, accounting for 53% of the variance. Thus, for ease of interpretation and to limit multiple testing, all 18 items were collapsed and treated as one overall measure of fake news judgments. Nonetheless, descriptive statistics for badge-level results are also presented in the Appendix.

### Attitudinal Certainty

Similarly, a single-item measure was presented alongside each of the news items, asking participants to indicate how confident they are in their reliability assessment on a 7-point Likert scale, ranging from not at all confident (1) to neutral (4) to very confident (7). Scale reliability analysis on the averaged 18 attitude certainty items (6\*3) indicated high internal validity (M = 5.23, SD = 0.84,  $\alpha =$  .89). Similarly, PCA results indicated that the items loaded on a single dimension with an eigenvalue of 3.88, accounting for 65% of the variance (see Appendix for badge-level results).

### Political Ideology

Political ideology was measured on a standard self-placement scale, ranging from 1 = very conservative, 4 = moderate, to 7 = very liberal. Although often more diverse than Mturk (Peer et al., 2017), the Prolific sample (M = 4.69, SD = 1.42) was fairly liberal with 21% conservatives, 18% moderates, and 61% identifying as liberal.

### 2.4 Results

A One-way ANOVA was conducted to compare the effect of treatment condition (inoculation, control) on the difference in pre-and-post reliability scores of the fake news items. Results demonstrate a significant main effect of treatment condition on aggregated reliability judgements: F(1, 196) = 17.54,

MSE = 0.36, p < .001,  $\eta 2 = .082)^3$ . Specifically, compared to the control condition, the shift in postpre difference scores was significantly more negative in the inoculation condition (M = -0.09 vs M = -0.45, Mdiff = -0.36, 95% CI [-0.19, -0.52], d = -0.60, Figure 2). A separate two-way ANOVA revealed no main effect F(2, 179) = 2.80, p = 0.06 nor interaction F(2, 179) = 0.96, p = 0.38 with political ideology<sup>4</sup>. In short, compared to their assessments on the pre-test, individuals demonstrated a larger decrease in perceived reliability of fake news items when in the inoculation group versus the control condition. Similar patterns were observed at the badge level in the game (see Appendix) although there was some heterogeneity across badges with average effect-sizes ranging from d = 0.14(polarization) to d = 0.58 (discrediting).



# Figure 2: Median difference (post-pre) in reliability assessments of fake news items across treatment conditions with jitter (Panel A) and density plots of the data distributions (Panel B).

Furthermore, a one-way ANOVA also demonstrated a significant main effect of treatment condition on (post-pre) confidence scores (Figure 3), F(1, 196) = 13.49, MSE = 0.27, p < .001,  $\eta 2 = .06$ . Mean difference comparisons across conditions indicate a significantly higher (positive) difference score in the inoculation group compared to the control condition (M = 0.22 vs. M = -0.06, Mdiff = 0.27, 95%

<sup>&</sup>lt;sup>3</sup> A linear regression with post-test as the dependent variable, condition as a dummy, and pre-test as a covariate gives the same result. There was no significant difference at pre-test between the conditions ( $M_{inoculation} = 3.14 \text{ vs. } M_{control} = 3.32$ ,  $M_{diff} = -0.185$ , 95% CI [-0.42, 0.005], p=0.12, see Appendix).

<sup>&</sup>lt;sup>4</sup> Though conservatives (M = 3.56) were significantly more susceptible than liberals (M = 3.05) on the pre-test, t(147) = 3.22, d = 0.61, p < 0.01, consistent with Roozenbeek and van der Linden (2019).

CI [0.13, 0.42], d = 0.52)<sup>5</sup>. This suggests that compared to their assessments prior to treatment exposure, individuals demonstrated a larger increase in confidence in the inoculation versus the control condition. Once again, a two-way ANOVA revealed no main effect F(2, 179) = 1.22, p = 0.30 nor interaction F(2, 179) = 0.14, p = 0.87 with political ideology. At the badge level (see appendix), effectsizes for increased confidence ranged from d = 0.23 (discrediting) to emotion (d = 0.49). Importantly, the increase in confidence only occurred for those (71%) who broadly updated their reliability judgments in the right direction<sup>6</sup> ( $M_{inoculation} = 0.29$  vs.  $M_{control} = -0.02 M_{diff} = 0.31$ , 95% [0.13, 0.49], t(126) = 3.37, p < 0.01). In contrast, no gain in confidence was found among those who either did not change or updated their judgments in the wrong direction ( $M_{inoculation} = 0.03$  vs.  $M_{control} = -0.11$ ,  $M_{diff}$ = 0.14 95% [-0.11, 0.39], t(68) = 1.13, p = 0.26).



Figure 3: Median change scores (post-pre) of confidence in reliability judgments across treatment conditions with jitter (Panel A) and density plots of the data distributions (Panel B).

<sup>6</sup> Meaning that fake headlines were deemed less reliable on the post-test compared to the pre-test (i.e.,  $M_{\text{diff}} < 0$ ).

<sup>&</sup>lt;sup>5</sup> linear regression with post-test as the dependent variable, condition as a dummy, and pre-test as a covariate gives the same result. There was no significant difference in confidence judgments at pre-test between conditions ( $M_{inoc}$ = 5.25,  $M_{control}$ =5.27,  $M_{diff}$ = 0.002, 95% CI[-0.24, 0.20]. *p* = 0.88, please see Appendix).

### 2.4 Discussion

This study successfully demonstrated the efficacy of a "broad-spectrum" inoculation against misinformation in the form of an online fake news game. Using a randomized design, multiple items, and measures of attitudinal certainty, this chapter replicated and expands on the initial study by Roozenbeek and van der Linden (2019). Overall, this study finds clear evidence in support of the intervention. Whereas Roozenbeek and van der Linden (2019) reported an average effect size of d = 0.52 for aggregated reliability judgments using a self-selected within-subject design, this study finds very similar effectsizes in a randomized controlled design (d = 0.60). The range in effect-sizes observed on the badge level (d = 0.14 to d = 0.58) are also similar to what Roozenbeek and van der Linden (2019) reported (d = 0.16 to d = 0.36) and can be considered sizeable in the context of resistance to persuasion research (Banas & Rains, 2010; Walter & Murphy, 2018). In fact, Funder and Ozer (2019) recommend describing these effects as medium to large and practically meaningful, especially considering the refutational-different rather than the refutational-same approach adopted here, i.e., in the game, participants were trained on different misleading headlines than they were tested on pre-and-post. The exposure to general manipulation techniques across different topics provides additional support for broad-scale inoculation interventions against misinformation. This phenomenon is consistent with recent research on the "blanket of protection", where the inoculation effect extends to previously unmentioned arguments on the same issue (Ivanov et al., 2012).

Importantly, consistent with Roozenbeek and van der Linden (2019), none of the main effects revealed an interaction with political ideology, suggesting that the intervention works as a "broad-spectrum" vaccine across the political spectrum. However, it is interesting that in both studies, the smallest effect is observed for the polarization badge. One potential explanation for the lower effect on polarization is confirmation bias: in the game, decisions can still be branched ideologically congenially. Given the worldview backfire effect (Lewandowsky et al., 2012), future research should evaluate to what extent inoculation is effective for ideologically congruent versus non-congruent fake news. Nonetheless, these results complement prior findings which suggest that susceptibility to fake news is the result of a lack of thinking rather than only partisan motivated reasoning (Pennycook & Rand, 2019). Lastly, the current study also significantly advances our understanding of the theoretical mechanisms on which the intervention acts. For example, while inoculated individuals improved in their reliability assessments of the fake news items, the average confidence they expressed in their judgements also increased significantly and substantially. Importantly, the intervention only

significantly increased confidence amongst those who updated their judgments in the right direction (i.e., correctly judging manipulative items to be less reliable).

### The case for therapeutic and generalised inoculations

These are promising findings in light of the limited contemporary understanding of the role of certainty in the inoculation process. By pushing the boundary conditions and assessing individuals' actual ability and confidence to differentiate between manipulative and non-manipulative content, these findings provide support for the efficacy of gamified, generalised, and therapeutic inoculation treatments. Furthermore, these results point towards a confidence-boosting effect of inoculation treatments on resistance. As pointed out by Tormala and Petty (2002), the longstanding assumption in inoculation research was that resistance equalled no attitude change. Here, the findings suggest that 'broad-spectrum' treatments against common manipulation techniques enhance inoculated individuals' ability to confidently and correctly spot and resist misinformation. In other words, this chapter provides initial findings arguing that certainty plays an integral part in the building and strengthening of attitudinal resistance, emphasising that something indeed *does* happen during the inoculation process. This is consistent with early theorising, where McGuire suggested that after attitude certainty was initially shaken by the threat component of the treatment message, the generation of and practice with counterarguments would increase confidence to resist (McGuire, 1964). Here, the gamified setup of Bad News which requires participants to actively raise and refute counterarguments themselves arguably offers an enhanced process of subvocal counterarguing which, in turn, seems to boost individuals' confidence in correctly resisting misinformation in the future.

### Situating attitude confidence in the inoculation process

Yet, a few deviations from traditional assessments of attitude certainty in persuasion are noteworthy. Firstly, the present study incorporated a meta-cognitive element of resistance by asking participants to reflect on their ability and confidence to resist misinformation. In other words, instead of assessing attitudes and attitude certainty before and after treatment exposure, the current study focuses on reliability assessments and confidence in judgements, respectively. This is consistent with Tormala and Petty (2002) who have emphasised the importance of a meta-cognitive account when examining resistance to persuasion. Specifically, in a series of four experiments, they found that when people believe that they have successfully resisted a persuasive message, they become more certain in their attitudes. However, this effect only occurred when the resisted attack is believed to be strong, allowing individuals to feel good about their meta-cognitive experience of resisting persuasion. On the other hand, when individuals perceived the resisted attack to be weak, the interference that their position on the issue topic is valid was not made as confidently. Hence, incorporating individuals' awareness of their confidence in resisting rather than towards a specific attitude, reflects an initial step towards rethinking the role of confidence in the resistance process. Regardless, more research is required to identify whether an increase in confidence pertains to the fake items themselves or rather the ability to refute misinformation in general. For example, Tormala and Petty (2004) have argued that these mechanisms are likely to be intertwined as individuals might be confident in their ability to refute counterarguments because they perceive their attitudes to be valid and therefore, are both more willing and likely to defend their beliefs. Additionally, research outlining the role of confidence in the spread of misinformation gives reason to surmise that confidence might play a bigger role in the inoculation process altogether by affecting not only the assessment of misinformation but also whether or not individuals choose to pass it on.

### Limitations and future research

As it is with all research, this study did suffer from a number of necessary limitations. First, the Prolific sample was likely not representative of the U.K. population. Indeed, a stark absence of applying inoculation cross-culturally is evident. If active inoculation is to be a scalable intervention against misinformation, its efficacy must be established across different cultures and socio-political contexts. Secondly, although the study controlled for modality (given that both Bad News and Tetris are games), it lacked a condition that is cognitively comparable to the inoculation condition. It will be important for future research to evaluate to what extent "active" gamified inoculation is superior to "passive" approaches—including traditional fact-checking and other critical thinking interventions—especially in terms of eliciting a) motivation, b) the ability to help people discern reliable from fake news, and c) the rate at which the inoculation effect decays over time. Third, although this study improved on the initial design by having participants evaluate simulated Twitter posts (pre and post) outside of the game environment, this study does not allow to determine if playing the Bad News game led to an increased ability to detect real news or changes in online behaviour (e.g., if players shared less fake news on social media than people who did not play the game). Fourth, the fact that a small minority of individuals appear to engage in contrary updating is worth noting and a finding future work may want to investigate further (e.g., in terms of prior motivations). Fifth, the duration and longevity of the inoculation effect was not assessed. Given that the inoculation effect is known to decay over time, albeit no clear parameters exist yet (Banas & Rains, 2010), future research should explore the durability of resistance through gamified inoculation treatments. Lastly, although the fictitious nature of the items helps rule out potential memory confounds and the lack of variation on the measures (prepost) in the control group should decrease concerns about potential demand characteristics, future research on decay should consider testing the reliability assessment of previously unseen items or examine whether the repeated assessment of the same content impacts the longevity of inoculation.

### 2.5 Conclusion

In conclusion, this study finds support for generalised and gamified inoculation treatments against the common manipulation techniques that underpin online misinformation. Simultaneously, the chapter

also addresses the main shortcomings evident in Roozenbeek and van der Linden's (2019) original evaluation of the Bad News game: the lack of a control group, a relatively small number of items to measure effectiveness, and the absence of attitudinal certainty measurements. Thus, the following can be concluded: compared to a control group, the generalized inoculation intervention not only successfully conferred resistance to online manipulation but also boosted confidence in inoculated individuals' ability to resist misinformation. Importantly, this confidence boost occurred in the correct direction and thereby, enhancing inoculated individuals' ability to *confidently* and *correctly* spot and resist misinformation. Though more research is needed, this research makes substantial steps towards understanding the mechanisms that build and strengthen resistance. Future research should pursue a more nuanced understanding of how attitude certainty may help enhance and extend the effects of inoculation-induced resistance against misinformation.

# 3 THE ROLE OF ATTITUDE CERTAINTY IN ONLINE SHARING INTENTIONS

### 3.1 Abstract

In light of the societal threat posed by fake news, recent research has explored the possibility to build psychological resistance to misinformation through inoculation. Inoculation theory is based on the biological analogy of an immunisation process, positing that pre-emptive exposure to weakened doses of manipulation motivates the development of "mental antibodies". This chapter aims to further explore the role of attitude certainty in the building and strengthening of attitudinal resistance. Having

previously established the efficacy of generalised inoculation treatments against misinformation on simulated social media platforms like Twitter, the current chapter aims to assess its efficacy within the neglected context of end-to-end messaging apps like WhatsApp. As part of a unique collaboration with WhatsApp Inc, we developed Join this Group, an online choice-based game that aims to inoculate people against the spread of misinformation on private messaging apps (e.g., peer pressure, trusted contacts). A randomised longitudinal study with a UK nationally representative sample (N=839) was conducted. Firstly, the results provide support for the notion that gamified prebunking interventions are an effective and scalable means of reducing susceptibility to misinformation encountered within the context of WhatsApp. Furthermore, a significant main effect of the intervention on reliability assessments of fake news items, attitude certainty, and willingness to share information online is evident. Building on Chapter 1, inoculated individuals also report being more confident in their assessments and less willing to share news items that employ manipulation strategies. Moreover, the results also suggest a mediating effect of confidence on sharing intentions of misinformation. Lastly, these findings are maintained for at least one week after playing Join this Group. These results have significant ramifications for designing misinformation interventions tailored to the specific challenges of encrypted messaging applications and raise additional questions regarding inoculation research's ability to spread resistance across issues, platforms, and individuals.

### 3.2 Introduction

### Misinformation protected by the walls of end-to-end encryptions

Over the years, social media companies have increasingly adopted and experimented with countermeasures against the prevalence of misinformation on their platforms (reference). Whether in form of flagging misleading content (Lanius et al., 2021) or taking down extremist groups (Ganesh & Bright, 2020; Gorwa et al., 2020), moderating content on social network platforms (e.g., Twitter, Facebook) is unlikely to stop the underlying psychological forces that lead people to believe in and share misinformation in the first place. Instead, content regulation appears to be a double-edged sword that can result in just as many unwanted by-products, including simply moving conversations to end-to-end encrypted messaging platforms such as WhatsApp, Signal, and Telegram (Badrinathan, 2021; Urman & Katz, 2020). Here, content can freely circulate within closed networks, creating a breeding ground for unverified, misleading, or false information (Garimella & Eckles, 2020; Resende et al., 2019). With over 2 billion WhatsApp users worldwide and a large

portion of news shared on the platform being false or distorted, WhatsApp is regarded as one of the biggest tools for spreading misinformation (Gross, 2017). Additionally, research shows that a significant amount of harmful and false content continues to be shared on these platforms, even after professional and third-party fact-checkers have debunked them (Reis et al., 2020). Research points out the role "closed" messaging applications play in undermining and disrupting political elections and referenda (Kazemi et al., 2021; Machado et al., 2019), spreading QAnon, a far-right conspiracy theory (Hoseini et al., 2021), as well as fuelling mob lynchings and the formation of "WhatsApp vigilante" groupings, that is, groups who take it upon themselves to enforce law and punishments in their neighbourhood (Arun, 2019b; Banaji, With, et al., 2019). This has led WhatsApp to take legal action against abusers of their platform (Kalra & Vengattil, 2019) and governments to urge WhatsApp to lift its encryption (Ellis-Peterse, 2021; Kazmin, 2018).

Despite WhatsApp's initial countermeasures, ranging from placing restrictions on forwarding messages to limiting the size of group chats, misinformation on the platform persists and was further exacerbated by the pandemic (Al-Zaman, 2021; Ferrara, 2020). Moreover, research suggests while such efforts can significantly *delay* the propagation of misinformation, they remain ineffective in *preventing* or stopping it (de Freitas Melo et al., 2020). Considering the previously mentioned shortcomings of algorithmic solutions and the inability to moderate content on private messaging applications, effective interventions that help individuals spot and resist misinformation are urgently needed. Yet, little is known about whether and how misinformation may differ on platforms that are closed, end-to-end encrypted, and commonly exclusively used to communicate with close social groups (e.g., family and friends). Given the more personal and private nature of such interactions, then, is it possible that misinformation spread on private messaging platforms is underpinned by different psychological factors?

### The psychological dynamics on private messaging platforms

A few characteristics unique to private messaging apps make the spread of misinformation particularly pervasive. At a psychological level, research on the effects of source information suggests that trusted endorsements (e.g., the sharing of content by a trustworthy source) significantly impact the perceived credibility of misleading content (Mena et al., 2020). Considering that the original source of online

information is often unknown or unfamiliar, a heightened reliance on social cues is evident when assessing message credibility (Jessen & Jørgensen, 2012; Seo et al., 2019). This emphasises the importance of social ties in the flow of information (Sun et al., 2006). Building on that, research underlines that more closely perceived connections exert different group peer pressures and encourage conformity (Bleize et al., 2021a; Brechwald & Prinstein, 2011).

Perhaps then, to mimic the settings unique to private messaging applications, a broader definition of misinformation is needed. Firstly, not all forms of misinformation are spread with ill intent. Basic human error, or put differently, beliefs that are genuinely believed to be true and therefore shared without any ulterior motives can be equally harmful (van der Linden, 2017). Additionally, social media is arguably not exclusively used to share one's convictions but to also seek out information and communities (Zhao & Zhang, 2017). Indeed, research on the psychology of rumours proposes that gossip arises in situations that are ambiguous or threatening and that it serves groups in their collective sense-making processes. Additionally, factors such as group protection, status enhancement, and feelings of belonging are assumed to be motivational drivers of engagement in *online* gossip (Cialdini & Trost, 1998; Lyons & Hughes, 2015; Talwar et al., 2019).

Consequently, one could argue that rumours and gossip spreading on private messaging platforms, especially in times of risks, uncertainties, and danger, represent a collective attempt at sense-making. However, more research is needed to enhance the current understanding of why people fall for and share misinformation in closed online conversations, and more importantly, whether attitudinal resistance can be built against such "psychological pitfalls" is crucial. To do so, scholars are increasingly paying attention to strategies that *pre-emptively* debunk ('pre-bunk'). In other words, rather than trying to undo harmful content, could we stop people from believing and sharing misinformation in the first place? Inoculation theory offers a promising step forward.

### *Inoculation theory – theoretical boundaries and extensions*

To briefly review, inoculation theory (McGuire, 1964; McGuire & Papageorgis, 1961; Papageorgis & McGuire, 1961) is based on the biological analogy of an immunisation process and posits that, just as injecting a weakened dose of a pathogen triggers the production of antibodies, pre-emptive exposure to persuasion motivates the generation of attitudinal resistance against future persuasion attempts. The theoretical pillars of such attitudinal immunisation consist of *threat* and *refutational pre-emption*. It is argued that invoking threat and an awareness of vulnerability to having one's attitudes attacked, motivates the generation of pre-emptive and attitude-boosting counterarguments against future persuasion attempts. Though the biological analogy has proven robust and efficient across a multitude of topics (for reviews and meta-analyses, see Banas & Rains, 2010; Compton et al., 2021; Lewandowsky & van der Linden, 2021), the recently renewed scholarly interest in inoculation theory has highlighted remaining gaps in the scientific understanding and has highlighted the need for

theoretical and practical revisions (Compton et al., 2021). For instance, while studies examining the decay of the inoculation effect have demonstrated effects of decay ranging from one week to 33 weeks (Pfau et al., 1992; Zerback et al., 2020), little is currently known about the longevity of inoculation treatments against online misinformation.

Recently, inoculation research has begun examining how inoculation theory may be situated within today's unique information infrastructure. Whether it is the differentiation between prophylactic and therapeutic inoculation approaches (inoculating the unexposed vs. already 'infected'), the generalisation of treatments (instead of issue-specific ones), or the pivot towards treatments that require *active* engagement (as opposed to traditionally passive treatments), novel avenues have emerged for inoculation scholarship (Compton et al., 2021). While these innovations are promising, several substantial gaps remain and need addressing for the application of inoculation to misinformation to be optimised.

### The Role of Attitude Certainty in Resisting Persuasion

While the previous chapter demonstrated a strong confidence-boosting effect of inoculation treatments on resistance against misinformation (Basol et al., 2020), a clear understanding of its role within the inoculation process itself is still absent. This is somewhat at odds with general persuasion scholarship, which has focused heavily on the effects of attitude certainty on attitudes and behaviours (see Tormala & Rucker, 2018, for a review). Specifically, findings suggest that attitudes held confidently are more likely to guide behaviour, help resist persuasion, and persist over time (Rucker & Petty, 2004; Tormala, 2015; Tormala & Petty, 2002). Yet, it remains mostly unclear as to *why* attitude certainty has such consequences. In other words, what is it about attitude certainty that facilitates and promotes the link between attitudes and behaviour as well as resistance to future persuasive arguments? And most relevant to this thesis, what role does it play in the inoculation process?

Reviewing existing, albeit more general, scientific discourse around the intricacies of attitude certainty, could provide opportunities to translate it into the mechanisms underpinning resistance through inoculation. Attitude certainty is regarded as a dimension of attitude strength (Petty & Krosnick, 1995) and refers to the degree of conviction that one's held attitudes are correct. Current conceptualisations of attitude certainty assume that the consequences of attitude certainty occur regardless of how certainty is reached or established. This is particularly interesting considering that having strong reasons behind one's attitude certainty seems to impact how easily one withstands persuasive attacks (Albarracín et al., 2000; Albarracín & Mitchell, 2004). This raises further questions and potential explanations as to how resistance is achieved. It becomes clear that, although the outcome of being resistant to persuasion can appear the same, individuals may have reached their resistance in different ways.

To give an example, people become more certain of their changed attitudes under high elaboration (Barden & Petty, 2008) while also becoming more certain of their initial attitudes when resisting persuasion under high elaboration (Tormala & Petty, 2004). On the contrary, research suggests that when people attribute their resistance to weak persuasive messages, they become less certain of their attitude than when it is perceived to withstand strong counterarguing (Tormala et al., 2006) Similarly, the perceived legitimacy of resistance seems to further impact how confidently attitudes are held(Tormala et al., 2007). Tormala and Petty (2002) suggest a more metacognitive framework, where when people believe that they have successfully resisted a persuasive message, they become more certain of their original attitude. These findings further emphasise that both the actual means by which attitudinal resistance is achieved as well as the mere perception of how this resistance was established have consequences on attitude certainty. Thus, it could be argued that these bases of one's attitude certainty can in turn affect the resistance against subsequent persuasive messages. Further understanding of when and how attitude certainty is established within the inoculation process will provide substantial insights into the process of inoculation as well as how resistance, in turn, may affect the conviction with which attitudes are held in the future. Consequently, it can be argued that the actual means by which individuals achieve resistance as well as their mere perceptions of how resistance was achieved have consequences on the degree of attitude certainty. Put differently, there is a possibility that the bases of one's attitude certainty—whether true or merely perceived—can have a different impact on the resistance against subsequent persuasive messages even when the degree of certainty is the same.

### Caring is sharing

Literature suggests that the benefits of attitude certainty can extend beyond attitudinal resistance. The perceived validity of an attitude can impact one's willingness to both defend and, more importantly, act in congruence with the attitude. Indeed, a growing body of research emphasises the impact of attitude certainty on attitude advocacy – making people more likely to talk about their attitudes, share their views, and even attempt to persuade others to adopt their views (Akhtar et al., 2013; Cheatham & Tormala, 2015; Visser et al., 2003). Importantly, this can be independent of the attitude itself and appears to be conceptually separate from attitude valence and extremity (whether and to what extent an attitude is positive/negative). In fact, Tormala and colleagues (2004) point out that the most germane question for current research is the notion that attitudes held with high confidence are more resistant than those held with little confidence (Bassili, 1996). Considering the fostering and guiding effect of attitude certainty on advocating behaviours of sharing and persuading, it remains unclear whether one's attitude certainty could have similar effects on online behaviour. Particularly in light of the mediating role of attitude certainty on resistance highlighted in Chapter 2, research should explore whether and how psychological inoculations can affect online sharing behaviour (Basol et al., 2020; Roozenbeek & van der Linden, 2019). Specifically, by exploring the possibility that the mediating role extends to the sharing of misinformation, this

chapter aims to examine the potential interplay between attitude confidence, attitudinal resistance, and sharing of misinformation. Establishing a clearer understanding of such mechanisms will allow future inoculation interventions to become more efficient reducing susceptibility to misinformation as well as the extent to which misinformation is shared and encountered in the real world in the first place.

### In Pursuit of a Solution with WhatsApp

With cases of violence and deaths fuelled by misinformation on WhatsApp rising (Arun, 2019; Banaji et al., 2019), we applied for and received the WhatsApp Research Grant against Misinformation<sup>7</sup>. We collaborated with WhatsApp's policy team to identify platform-specific challenges and developed, *Join this Group<sup>8</sup>*, a gamified inoculation treatment against misinformation. Having provided initial support for gamified, generalised, and active inoculation treatments, this collaboration facilitated the study of critical factors which influence susceptibility to and the sharing of misinformation. future interventions can and should pivot towards more ecologically rich set-ups that incorporate the critical factors which influence susceptibility to and the sharing of misinformation. Specifically, by simulating the infrastructure and context of private-messaging apps and incorporating elements such as group dynamics, normative influences, 'social etiquette', and forwarding options, this project tests the efficacy of an inoculation treatment under 'noisier' and ecologically richer circumstances. In short, *Join This Group* represents the effort to incorporate the insights gained in Chapter 2 into an intervention that allows to further manipulate, dissect, and leverage the mechanisms underpinning inoculation-induced resistance.

Yet, similar to vaccination roll-out campaigns, the mere presence of an inoculation treatment is not sufficient. Rather, vaccination campaigns account for the complex interplay of differing factors. Similarly, to obtain psychological "herd immunity" against misinformation, treatments must be desired, accessible, safe, and effective against the many 'variants' of misinformation. Therefore, the "misinfodemic" (Yasmin, 2021) necessitates a treatment which is not conceptually bound or restrained

<sup>&</sup>lt;sup>7</sup> This grant and project were jointly received and conducted with my collaborators Dr Jon Roozenbeek and Prof van der Lidnen; https://www.poynter.org/fact-checking/2018/whatsapp-awards-1-million-for-misinformation-research/

<sup>&</sup>lt;sup>8</sup> https://whatsapp.aboutbadnews.com/#/intro

to a certain modality. In other words, a treatment condition against misinformation should ideally be effective against a variety of topics emerging in a variety of forms. Rather than inoculating against specific arguments within a particular topic then, *generalised* and *therapeutic* inoculation treatments can confer resistance against related yet untreated topics *and* for those who have previously been exposed to misinformation. Prioritising the scalability and adaptiveness to platform-specific challenges are essential building blocks to achieve psychological 'herd immunity against misinformation (Compton et al., 2021). Lastly, research has yet to establish the prerequisites and boundary conditions of spreading inoculation from one person to another. Whether psychological 'herd immunity requires each individual to receive an inoculation treatment or whether the inoculated could pass on attitudinal resistance to others (via '*vicarious inoculation*') remains an open question. To summarise then, this chapter aims to further examine attitude certainty and its role in the spread of misinformation as well as in *building* and *strengthening* resistance conferred through inoculation.

### **PRESENT STUDY**

This chapter aims to continue contributing towards a more nuanced portrayal of certainty in the inoculation process. Is attitude certainty a prerequisite or an unintentional yet beneficial by-product of inoculation treatments? Considering research that emphasised the role of attitude certainty in inoculation-induced advocacy, the natural next question, then, concerns itself with whether the confidence-boosting effect noted in the Chapter 2 extends itself to behaviours that underlie the sharing of online (mis)information, especially within the context and challenges of private messaging platforms. To understand whether and how attitude certainty can enhance and spread resistance, the present study examined attitude certainty and its potential role in sharing behaviours to the context of online misinformation on end-to-end encrypted messaging apps.

### 3.3 Developing Join this Group

Following the inoculation metaphor, the game exposes individuals to weakened forms of misinformation by *actively* letting them generate their own manipulative content in a fictitious environment simulating a private messaging platform. However, rather than following the traditional issue-specific set-up of the inoculation treatment, the presented active inoculation intervention aimed to inoculate against the very tactics that underlie the production and spread of misinformation (i.e., a *generalised* and analogous to a broad-spectrum vaccine). While there is growing evidence for the relative benefits of "active" inoculation treatments, previous gamified interventions such as the *Bad News* game and *Harmony Square* (Roozenbeek & van der Linden, 2020) have exclusively focused on misinformation on public social media platforms (such as Facebook and Twitter). This reduces the

potential applicability of these games in countries where direct messaging apps play a more dominant role in communications and information-seeking behaviour.

Thus, in Join this group, individuals earn "badges" that correspond with manipulation techniques commonly present in misinformation on direct messaging apps, namely, impersonating an expert (Goga, Venkatadri, et al., 2015; Jung, 2011; Reznik, 2012) using emotional language to frame and misconstrue content (Konijn, 2013; Zollo et al., 2015), polarising narratives to elicit hostility towards out-group (Groenendyk, 2018; Iyengar & Krupenkin, 2018) and escalating an issue such that misinformation informs offline behaviour and causes acts of aggression (Banaji et al., 2019b). These theory-driven strategies are partially derived from NATO's report "Digital Hydra", which outlines the various forms of misinformation strategies and was also based on information from WhatsApp to ensure the intervention has ecological validity. The notion behind an *active* "psychological vaccine", then, is to let individuals generate their own "antibodies". Join this Group incorporates the inoculation components by utilising 1.) warnings about imminent fake news and 2.) pre-exposure to weakened doses of manipulation tactics in form of fictitious content where, instead of leading players to spread fake news, individuals "learn by doing". Research suggests that both processes enhance the inoculation effect by facilitating retention in memory (Pfau et al., 2005, 2006). Indeed, a large literature exists on the benefits of simulations and games in achieving educational outcomes (for systematic review, see (Boyle et al., 2012) Specifically, Przybylski and colleagues (2010) explain that games enhance motivation by letting individuals immerse themselves in a virtual identity and tap into basic psychological needs of competence (understanding, learning, goals, challenges), autonomy (flexibility to choose, create your own path) and relatedness (feedback, interaction). Lastly, adapting a follow-up design will provide additional insights into the longevity and decay of the inoculation effect induced by gamified inoculation treatments.

### 3.4 Methods

### Sample

To obtain a national sample of the UK, participants were recruited via *Prolific Academic* (Peer et al., 2017). Participants who completed the full study (including 2-3 min follow-up study) received £2.25 in compensation. Taking into account the average inoculation effect reported in previous research (Roozenbeek & van der Linden, 2019), an a priori power analysis was conducted with G\* power using  $\alpha = 0.05$ , f = 0.25 (d = 0.52) and power of 0.95 with two experimental conditions and two repeated measures. The minimal sample size required for detecting the main effect was n= 158. A nationally balanced (age, gender) UK sample (N=923) was recruited, and inclusion criteria were age, fluency in English, and usage of WhatsApp. In total, 839 participants took part in the one-week follow-up study (9% attrition). 52.4% of our sample identified as female (47.4% female, 0.1% other); 55.2% indicated

being between 25 and 44 years of age, and 34.7% reported having a university bachelor's degree. Our sample also skewed politically left (M = 3.45, SD = 1.39).

### Measures

With each item during the item-rating task (pre and post), participants completed three questions about the perceived reliability of the item, their confidence in the reliability assessment, and their willingness to share the item on their social media. These measures were specifically created for the context of WhatsApp and the purpose of this study.

### Reliability

To assess participants' perceived reliability, a single-item measure was presented alongside 12 (4\*3) fake items that looked like screenshots of WhatsApp chats and 4 real news items (1\*4) (example item polarization; "Check this! New interview FAIL. Crime is on the rise like crazy and (blanked out)'s solution is to "listen" to what criminals want! Insane!! ", see Figure 4 for examples). Out of these 12 fake items, 3 items corresponded with one of 4 badges (fake experts, emotional language, polarisation, escalation) that participants earned in Join this Group (See Appendix for all items). Participants reported the perceived reliability of the shared content in the group chat on a 7-point Likert-scale from not reliable at all (1), neutral (4) to very reliable (7). Following Roozenbeek and van der Linden (2019), to form a general fake news scale of perceived reliability of fake items, all 12 fake news items were averaged. An initial reliability analysis suggested good internal consistency (M = 2.67, SD =0.99,  $\alpha = 0.87$ ) of the 12-item fake news scale. A subsequent exploratory principal component analysis (PCA) was also run on the fake news items. According to the Kaiser criterion, results indicated that the items clearly loaded on a single dimension with an eigenvalue of 5.08, accounting for 42.3% of the variance. Thus, for ease of interpretation and to limit multiple testing, all 12 items were collapsed and treated as one overall measure of fake news judgments.



Figure 4: Screenshots simulating WhatsApp conversations as fake news items (polarisation on the left, escalation on the right).

### Attitude Certainty

Similarly, a single-item measure was presented alongside each of the fake news items, asking participants to indicate how confident they are in their reliability assessment on a 7-point Likert scale, ranging from not at all confident (1) to neutral (4) to very confident (7). Scale reliability analysis on the averaged 12 attitude certainty items (4\*3) indicated high internal validity (M = 5.4, SD = 1.03,  $\alpha$  \_= .93). Similarly, PCA results indicated that the items loaded on a single dimension with an eigenvalue of 6.93, accounting for 57.79% of variance.

### Sharing

Lastly, a single-item measure was presented alongside each of the news items, asking participants how likely they are to forward the message to others on a 7-point Liker scale, ranging from not at all (1) to neutral (4) to very likely (7). Scale reliability analysis on the averaged 12 willingness to share fake items (4\*3) indicated high internal validity (M = 1.83, SD = 0.96,  $\alpha_{-} = .91$ ). Similarly, PCA results indicated that the items loaded on a single dimension with an eigenvalue of 6.17, accounting for 51.48% of the variance.

### Procedure

This study employed a 2 (*Join this group* vs. Control) x 2 (pre-post) mixed design to test the efficacy of gamified inoculation interventions in conferring attitudinal resistance against platform-specific misinformation. Participants were randomly assigned to one of two conditions (inoculation, control). The inoculation condition entailed playing the WhatsApp game ('*Join this group*'), whereas participants in the control condition were asked to play Tetris for approximately 10 minutes, the same amount it takes to play *Join this group*. We chose *Tetris* for several reasons: (1) it has successfully been used in previous studies examining gamified inoculation (Basol et al., 2020; Maertens et al., 2020); (2) it is publicly available; (3) and it is a simple game with a flat learning curve.

To begin with, participants performed an item-rating task where they were randomly shown 16 items (4 real, 12 fake items). As previously described, four theory-based common manipulation techniques (fake expert, emotional language, polarisation, and escalation) served as the basis for the 12 fake items (3 items/strategy). Alongside each item, participants were also asked to rate the 16 items on a 1-7 scale (1 being 'Not at all' and 7 being 'Very"): (1) How reliable do you find this post? (Roozenbeek & van der Linden, 2019); (2) How confident are you in your judgement? (attitudinal certainty; Basol et al., 2020); (3) How likely are you to forward this message to others? (Roozenbeek & van der Linden, 2020). Consistent with previous studies, source information was blacked out to avoid source confounds (Roozenbeek & van der Linden, 2020).

Upon completion of this item rating task, participants were randomly assigned to one of the treatment conditions (inoculation or control). The inoculation condition required participants to play through *Join this Group* (see Figure 5), where throughout four separate fictitious scenarios set in an environment simulating WhatsApp, players learned about and how to make use of the four manipulation techniques to spread misinformation (fake experts, emotional language, polarisation, escalation). Additionally, what differentiates this game from both *Bad News* (Chapter 1) and *Go Viral!* (Chapter 3) is the incorporation of group-based peer pressures and group dynamics. Thus, consistent with research illustrating the cyber aggression and conformity on WhatsApp (Aizenkot & Kashy-Rosenbaum, 2018; Bleize et al., 2021a; Brechwald & Prinstein, 2011), these scenarios increasingly got more extreme and 'explosive' (e.g., it starts with spreading health misinformation about kiwis causing cancer to instigate a riot). Participants follow the choice-based structure of the game and collect points for each decision. When players make poor or wrong decisions that do not employ the misinformation technique well, they "get banned" from the group. Participants are told that the game will be over once they receive three bans.



Figure 5: Join this Group landing page (left) and game environment (middle and right).

Additionally, with each scenario, players receive a 'badge' for the learned manipulation technique. To ensure that all participants in the inoculation condition played attentively, a password, which was displayed at the end of the game, was required to have their submission accepted. Equally, participants who demonstrated low-effort responses (e.g., same rating for all items) were excluded and resampled. Subsequently, participants were asked to rate the reliability of the items, their confidence in their reliability assessment, and their likelihood of forwarding the post. After completing a series of demographic questions (e.g., age, gender, political ideology), participants were debriefed and reminded that they would be recontacted a week later for the follow-up study. Accordingly, participants received an invitation to partake in the follow-up survey a week later, where they completed the same item rating task (with perceived reliability, confidence, and willingness to share as outcome measures) for the same 16 items (4 real and 12 misinformation/ 3 per technique posts learned in *Join this group*). See Figure 6 for the study design. This study was approved by the Cambridge University Ethics committee (REC-2018-19/19).



### Figure 6: Study design flowchart.

For this study, the following hypotheses were tested:

 $H_1$ : Participants in the *Join this group* condition will assess the real and misinformation items more accurately than the control condition.

 $H_2$ : Participants in the inoculation condition will be more confident in their reliability assessments than the control condition.

 $H_3$ : Participants in the inoculation condition will be less willing to share misinformation with others in their network than the control condition.

 $H_4$ : One week after exposure to the intervention, participants in the inoculation conditions will display minimal decay of the inoculation effect (for reliability, confidence, and sharing).

H<sub>5</sub>: Heightened attitude certainty will mediate willingness to share misinformation with others.

### 3.5 Results

Firstly, the analyses for the three main outcome measures included in the item rating task (reliability, confidence, and sharing) will be presented separately, for both the misinformation and real items, focusing primarily on the *difference* (post-pre) for each outcome measure before and after the treatment between both conditions<sup>9</sup>.

### Reliability

A one-way ANOVA<sup>10</sup> shows a significant effect of condition (*Join this group*, Control) on the prepost difference in the perceived reliability of misinformation items presented in fictitious conversations on WhatsApp (F(1,837) = 101, p < 0.001,  $\eta^2 = 0.108$ ). A Tukey HSD post-hoc comparison shows that the pre–post difference in perceived reliability for the *Join this group* condition was significantly higher than the control condition (M = -0.62 vs M = -0.2,  $M_{diff} = 0.42$ , 95% CI (0.34–0.5),  $p_{tukey} < 0.001$ , d = 0.69). Hence, playing the *Join this group* game significantly decreases the perceived reliability of misinformation encountered in WhatsApp chats (see Table S6 for item-level statistics). Figure 7 shows these results in a violin plot.

<sup>&</sup>lt;sup>9</sup> There were no mean pre-test differences between conditions for reliability (p=0.24), confidence (p=0.13), nor sharing intentions (p=0.35).

<sup>&</sup>lt;sup>10</sup> No mean pre-test differences between treatment condition and the control condition are apparent.



Figure 7: Pre-post differences in reliability scores of fake news items between conditions.

For real news items, we also find a significant effect of condition on the pre-post difference in perceived reliability (F(1,837) = 20.3, p < 0.001,  $\eta^2 = 0.024$ ). A Tukey post-hoc comparison shows that the real news was perceived as significantly less reliable in the Join this group condition than in the control condition ( $M_{inoc} = -0.25$  vs  $M_{control} = -0.05$ ,  $M_{diff} = 0.2$ , 95% CI (0.116-0.295),  $p_{tukey} < 0.001$ , d=0.31). To test whether people who played Join this Group were significantly more accurate in distinguishing between real news and misinformation, a paired sample t-test on the pre-and postgameplay difference for the difference in reliability scores between misinformation and real news was conducted (i.e., the level of 'veracity discernment'). Doing so, gives a significant post-gameplay increase in veracity discernment, showing that Join this Group players are better able to differentiate real news and misinformation after gameplay, ( $M_{\text{discernment,pre}} = 0.31$ ,  $M_{\text{discernment,post}} = -.0441$ ,  $M_{\text{diff}}$ =0.36, 95% CI (0.28- 0.43), t(2,379)=9.31, p<0.001, d=0.58, 95% CI (0.37-0.58). Thus, these findings provide partial support for hypothesis  $H_1$ : playing Join this group initially decreases the perceived reliability of misinformation presented within the context of WhatsApp but also of real news (although the effect size is about 50% smaller than for misinformation) and inoculated individuals demonstrated higher levels of truth discernment when differentiating between true and false news. Finally, a linear regression was run with the pre-post difference scores for perceived reliability as the dependent variable, and age, gender, educational attainment, and political ideology as covariates. Results suggest no significant effects (all ps > 0.08), except for political ideology (p = 0.04), so that identifying as left-wing is associated with a higher post-pre inoculation effect in terms of reliability assessment than people who identify as right-wing (see Appendix).

### Follow-up

To test whether this observed effect changed over time, a repeated measures one-way ANOVA was conducted with condition (*Join this group vs.* control) as the between-subject factor, and time (pre - post - follow-up) as the within-subject factor. Mauchly's Test of Sphericity indicated that the assumption of sphericity was violated,  $\chi^2(2) = 0.921$ , p < .001 and therefore, a Greenhouse-Geisser

correction was used. Doing so, illustrates a significant effect of time x condition on the perceived reliability of misinformation, F(1.85, 1551.8) = 168.8 p < .001,  $\eta^2 = .028$ . Specifically, a week later, participants in the control condition rated misinformation as significantly more reliable than inoculated individuals, ( $M_{control} = 2.53 \text{ vs } M_{inoc} = 2.03$ ,  $M_{diff} = 0.49$ , 95% CI (0.32, 0.66),  $p_{tukey} < 0.001$ , d = 0.39). There is a significant main effect of intervention on the average inoculation effect, F(1, 837) = 4.67, p = .003,  $\eta^2 = .006$ ). A difference-in-difference analysis ( $M_{diffT3T2,control} = 0.004$ ,  $SD_{diffT3T2,control} = 0.75$ ;  $M_{diffT3T2,inoc} = 0.11$ ,  $SD_{diffT3T2,inoc} = 0.66$ ) using the Tukey post-hoc test, indicates a significant mean difference of  $M_{diff-diff} = -0.107$ , t (837) = -2.16,  $p_{tukey} = 0.03$ , 95% CI (-0.2, -0.01), d = -0.15. These results provide partial support for **H**<sub>4</sub>: demonstrating that though there is a decay effect, inoculated individuals rated fake items as significantly less reliable than participants in the control group one week after treatment exposure (see Figure 8).



Figure 8: Between conditions difference in the perceived reliability of fake items over time.

### Confidence

For the confidence measure, a between-subjects ANOVA on the pre–post difference in confidence scores for misinformation is significant F(1,837) = 48.6, p < 0.001,  $\eta^2 = 0.055$ ), in that participants in the inoculation condition are significantly more confident after the intervention in their reliability assessment of misinformation than the control group ( $M_{control} = -0.07$  vs  $M_{inoc} = -0.43$ ,  $M_{diff} = 0.35$ , 95% CI (0.25, 0.45), p < 0.001, d = 0.5). Similarly, for real news, a between-subjects ANOVA shows a significant difference between conditions for the pre–post difference in confidence scores (F(1,837) = 6.14, p < .001,  $\eta^2 = 0.007$ ). A post-hoc comparison shows that the pre-post difference in attitude confidence is significantly higher in the *Join this group* condition than the control condition ( $M_{control} = -0.001$  vs  $M_{inoc} = -0.13$ ,  $M_{diff} = -0.13$ , 95% CI (0.03, 0.31),  $p_{tukey} = 0.01$ , d = 0.17). Thus,

participants in the inoculation condition reported significantly higher levels of confidence in the posttest than the control condition. Hence, these results support  $H_2$ : by demonstrating that inoculated individuals are more confident in their reliability assessments than those in the control group.

Additionally, to examine whether participants became more confident in their reliability assessments if they also *correctly* perceived the fake items as less reliable, an ANOVA was conducted. Here, the pre-post difference in misinformation confidence served as the dependent variable, while condition and "updated reliability" (a binary variable that is positive if the pre-post reliability difference score for fake items is positive and negative if this difference is negative) as fixed factors. Doing so shows a significant effect of condition x updated reliability on misinformation confidence, F(2,833) = 4.50, p = 0.01,  $\eta^2 = 0.010$ ). More specifically, post-hoc comparisons show that there is a significant difference in condition on correct confidence boosts ( $M_{control} = -0.48$  vs  $M_{inoc} = -0.75$ ,  $M_{diff} = 0.27$ , 95% CI (0.28, 0.64),  $p_{tukey} < 0.001$ , d = 0.46). The "confidence boost" only occurred in the right direction, meaning that inoculated individuals who *correctly* assessed fake items as less reliable in the post-test did so more *confidently*. See Figure 9 for a breakdown of the confidence-boosting effect on the difference in reliability assessments.



A higher difference represents a larger negative change in reliability assessments

## Figure 9: Between condition differences in updated confidence (pre-post) in reliability assessments of fake items.

To test whether this observed effect changed over time, a repeated measures one-way ANOVA was conducted with condition (*Join this group*, control) as the between-subject factor, and time (pre - post - follow-up) as the within-subject factor. Mauchly's Test of Sphericity indicated that the assumption of sphericity was violated,  $\chi^2(2) = 0.87$ , p < .001 and therefore, a Greenhouse-Geisser correction was

used. Doing so demonstrates a significant main effect of intervention on the confidence in reliability assessments F(1, 837) = 13.0, p = .001,  $\eta^2 = .012$ ). Furthermore, the results suggest a significant effect of time x condition on individuals' confidence in their judgement of misinformation, F(1.77, 1485.06)= 19.6, p < .001,  $\eta^2 = .004$ . Additional difference-in-difference analysis ( $M_{diffT3T2,control} = -0.01$ ,  $SD_{diffT3T2,control} = 0.95$ ;  $M_{diffT3T2,inoc} = 0.01$ ,  $SD_{diffT3T2,inoc} = 0.95$ ) using a post-hoc *t*-test indicates a nonsignificant mean difference of  $M_{diff-diff} = -0.02$ , t (837) = -0.41, p = 0.68, 95% CI (-0.15, 0.1), d = 0.02. Consistent with **H**<sub>4</sub>: a week after treatment exposure, no decay effect is evident and inoculated participants remain significantly more confident in their reliability assessments compared to the control group (see Figure 10).



Figure 10: Confidence scores between conditions throughout time points (pre, post, follow-up).

### Sharing

A one-way ANOVA shows a significant effect of condition (*Join this group*, Control) on the pre-post difference in willingness to share misinformation items presented in a stimulated WhatsApp conversation (F(1,837) = 14.03, p < 0.001,  $\eta^2 = 0.016$ ). A Tukey HSD post-hoc comparison shows that the pre–post difference in willingness to share for the control condition was significantly higher for the control condition than for the *Join this group* condition ( $M_{control} = -0.09$  vs  $M_{inoc} = -0.22$ ,  $M_{diff} = 0.12$ , 95% CI (0.05, 0.18),  $p_{tukey} < 0.001$ , d = 0.26). Hence, playing the *Join this group game* 

significantly decreased the reported likelihood of sharing misinformation encountered on WhatsApp<sup>11</sup>. Fore real items, the analysis also finds a significant effect of condition on the pre–post difference in willingness to share (F(1,837) = 6.23, p = 0.01,  $\eta^2 = 0.007$ ). A Tukey post-hoc comparison shows that the willingness to share real news is significantly lower in the *Join this Group* condition than in the control condition ( $M_{control} = -0.059$  vs  $M_{inoc} = -0.152$ ,  $M_{diff} = 0.093$ , 95% CI (0.116–0.037),  $p_{tukey} = 0.01$ , d = 0.17). Thus, these findings provide partial support for hypothesis **H**<sub>3</sub>: compared to the control condition, playing *Join this Group* initially decreases the willingness to share misinformation items presented within the context of WhatsApp but also of real news (although the effect size is descriptively smaller than for misinformation).



Figure 11: Difference scores in reliability, confidence, and sharing of fake news items across conditions.

### Follow-up

A repeated measures one-way ANOVA was conducted with condition (inoculation, control) as the between-subject factor, and time (pre - post - follow-up) as the within-subject factor. Mauchly's Test

<sup>&</sup>lt;sup>11</sup> See Appendix for item-level descriptive statistics.

of Sphericity indicated that the assumption of sphericity was violated,  $\chi^2(2) = 0.87$ , p < .001 and therefore, a Greenhouse-Geisser correction was used. There is a significant main effect of intervention on the willingness to spread fake items F(1, 837) = 7.35, p = .007,  $\eta^2 = .008$ ). Furthermore, results illustrate a significant effect of time x condition on the participants' willingness to share misinformation, F(1.78, 1490.58) = 5.33, p < .0007,  $\eta^2 = .001$ . Additional difference-in-difference analysis ( $M_{diffT3T2,control} = 0.007$ ,  $SD_{diffT3T2,control} = 0.60$ ;  $M_{diffT3T2,inoc} = 0.03$ ,  $SD_{diffT3T2,inoc} = 0.62$ ) using a post-hoc *t* test indicates a non- significant mean difference of  $M_{diff-diff} = -0.018$ , t (837) = -0.44, p = 0.65, 95% CI (-0.1, 0.06), d = 0.03. This supports H<sub>4</sub>: in showing that the inoculation effect on sharing does not decay after a week. See Figure 12 for visualisation of willingness to share misinformation items over time.



Figure 12: Sharing of fake items over time (pre, post, follow-up) between conditions.

### Mediations

Next, a mediation analysis was run to examine whether confidence played a mediating role in the sharing intentions. The independent variable consisted of the experimental condition (control, inoculation), the outcome variable was the willingness to share news (fake/real items), and the mediator variable for the analysis was confidence. The results revealed that the total effect of the treatment conditions on sharing misinformation was significant,  $\beta = -0.12$ , t = -3.75, p < .001, 95% CI(-0.18, -0.05). With the inclusion of the mediating variable (confidence in assessment of fake items), the impact of condition on sharing was still found significant,  $\beta = -.09$ , t = -2.97, p = .003, CI(-0.16, -0.02). The indirect effect of the treatment conditions on sharing of sharing of misinformation items is partially mediated (18.8%) by participants' confidence in their reliability assessment (see Figure 13). Thus, we find support for **H**<sub>5</sub>: attitude

certainty mediated participants' willingness to share misinformation with others. On the other hand, the results revealed a significant total effect of treatment on the sharing of real news  $\beta = -0.17$ , t = -2.5, p=0.012, CI (-0.3, -0.03). Although the direct impact of condition on confidence was significant,  $\beta = -0.16$ , t = -2.3, p = 0.02, CI(-0.31, -0.02), the indirect effect is insignificant,  $\beta = -0.012$ , t = -1.6 p = 0.1, CI(-0.042, -0.001) suggesting that there is no mediating effect of confidence on sharing of real news.



Figure 13: Path plot of mediation analysis on the relationship between condition and sharing intentions as mediated by confidence. \*p <.05.

### 3.6 Discussion

### Gamified Inoculation across topics, platforms, and individuals

The present studies primarily aimed to investigate the efficacy of active, generalised, and therapeutic inoculation interventions against misinformation encountered on private messaging applications. The findings are consistent with the previous chapter by finding support for the role of confidence as a mechanism for the effectiveness of inoculation, , as well as the reduction of misinformation sharing (Basol et al., 2020). These findings are also supported by previous research demonstrating that as people become more confident in their attitudes, they also become more willing to speak about them and even persuade others to adopt their views (Akhtar et al., 2013; Visser et al., 2003). In addition to this, this chapter provides support for gamified inoculation intervention against online misinformation about a variety of topics. Having reviewed two separate games (*Bad News, Join this Group*) which each concerned themselves with different topics and forms of misinformation, the accumulated findings further underline the efficacy of active inoculation efforts spanning across topics, platforms, and potentially individuals. This chapter finds that inoculated individuals perceive misinformation items as less reliable, are more confident in their judgements, and less willing to share falsehoods with others. Importantly, this increase in confidence only occurs in the right direction. In other words, only individuals became more confident in correctly rating real and fake news items.

Additionally, the findings are consistent with and build up on Chapter 2 by demonstrating the mediating role of attitude certainty on the spread of misinformation. However, this finding does not apply to real news items and given that *Join this Group* predominantly sheds light on manipulation techniques (rather than what makes information trustworthy and credible), it can be argued that one would not expect a mediating role of attitude confidence for real items. Lastly, though there was a mild decay effect on perceived reliability a week after treatment exposure, inoculated individuals still assessed misinformation as significantly less reliable than the control condition. Additional insights into the longevity of the inoculation effect including attitude certainty and sharing intentions contribute to the current scientific understanding of decay (Maertens et al., 2020; Compton, 2013). This is particularly important given research which demonstrated that while debunking efforts, such as fact-checking, decrease the likelihood of sharing falsehoods, these effects do not consistently and reliably persist in the long run for them to be employed as the only effort against online misinformation (Friggeri et al., 2014).

# Generating, strengthening, and spreading inoculation through attitude certainty

Building upon previous research demonstrating the confidence-boosting effect of gamified inoculation interventions (Basol et al., 2020, 2021), both studies contribute to the current understanding of the role of confidence in the inoculation process in multiple ways. Firstly, the mediating effect of the inoculation-induced confidence boost on the reliability assessment of fake news items builds onto the confidence boost reported in Chapter 2. Additionally, the mediating effect of confidence on sharing intentions of misinformation extends these findings and suggests that on top of confidence playing a key role in the generation and strengthening of resistance, it may also play a crucial role in the spreading of inoculation itself. The heightened confidence in one's ability to correctly assess the reliability of misinformation appears to influence their engagement with manipulative content. This is a promising step toward a "psychologically inoculated" individual who will detect, resist, and not pass on manipulative content. Additionally, attitude confidence does not appear to have a mediating effect on the sharing of real news items, suggesting that the game successfully targets one's ability to spot and interrupt the flow of false information specifically. This is particularly promising considering the findings that showed that playing Join this Group initially decreased the willingness to share misinformation, but also real items presented within the context of WhatsApp (although the effect size is descriptively smaller than for misinformation). This sheds some additional light on how attitude confidence may operate within the inoculation process, functioning as a fine-tuning mechanism that allows resistance to be more nuanced. In other words, inoculated individuals experience a confidence boost in their reliability assessments of fake items and report being less willing to spread fake and real items.

Yet, when looked at together, it becomes clear that attitude certainty plays a role in what inoculated individuals intend to share online. Given that confidence was directly manipulated by inducing a boost in the treatment conditions and not in the control condition, the findings provide strong support for the notion that attitude certainty plays an important mediating part in the inoculation process as well as in the behaviours after-treatment exposure (i.e., sharing misinformation). The observed mediating effect of attitude confidence on sharing intentions can be understood within the context of existing research. To give an example, it can be hypothesised that inoculation treatments disproportionately impact the subjective correctness and validity of an attitude (attitude correctness) over how confidently one is aware of their attitudes (attitude clarity) (Petrocelli & Rucker, 2007). Additionally, considering research (Holland et al., 2003) emphasising the strengthening effect of frequently expressed attitudes on attitude certainty, future research should explore whether frequency and depth of sharing inoculation-congruent content (whether due to advocacy or reassurance), in turn, impacts confidence and resistance (Compton & Pfau, 2009; Dillingham & Ivanov, 2016). Similarly, future research should also investigate whether a more meta-cognitive approach to perceived resistance affects inoculationinduced attitude certainty. That is, whether the perceived difficulty of resisting and counterarguing attack messages has any impact on resistance and the sharing of misinformation (Tormala & Petty, 2002).

These findings are consistent with research outlining the role of attitude certainty in resisting persuasion. Additionally, results presented in this chapter suggest that attitude certainty plays a crucial role in when and why people resist and disrupt the prevalence of harmful information. To complete that picture, however, future research should investigate the potentially beneficial effect of sharing inoculation through talk on attitude certainty and vice versa. Disentangling these processes and understanding how individuals can become more correctly confident while also enhancing the likelihood of spreading inoculation across social networks constitutes a crucial step towards combating the threat of misinformation. Additionally, such findings raise a new set of questions regarding the potential of inoculation research to combat misinformation. To contribute to our current scientific understanding, however, it is crucial to identify what remaining questions this research can and cannot answer. Thus, it is necessary to reflect on some of the shortcomings and highlight where more research is needed to utilise inoculation treatments to their full potential.

### Limitations and remaining questions

Firstly, the data is self-reported and therefore, does not provide insights into how *Join this Group actually* impacts real-world behaviour. Secondly, the items tested in the follow-up study remained identical to the headlines presented in the pre-post-tests, potentially functioning as "booster shots" or confounding the results as simple testing effects. In other words, repeated exposure to manipulation techniques could function as a reinstatement of the inoculation effect (Ivanov, Parker, et al., 2018;

Maertens et al., 2020). This is consistent with a wealth of cognitive research, demonstrating the memory-strengthening effect of repeated testing (Karpicke & Roediger, 2008). Future research should therefore test the long-term efficacy of gamified inoculation interventions on spotting and resisting *previously unseen* forms of manipulation.

Undeniably, more research is needed to establish a more nuanced understanding of the mechanisms and the dynamic role of attitude certainty within the inoculation process as well as its impact on online sharing behaviour. Some researchers have argued that the sharing of false content should not be taken as evidence for widespread false beliefs and that instead, veracity has little impact on sharing intentions. Or rather, that people are likely to share information they *could have* identified as inaccurate (Pennycook et al., 2021; Pennycook & Rand, 2020). Although *Join this Group* mimicked the interface of WhatsApp and aimed to explore the efficacy of inoculation within the specific context of private group chats, the fact that the messages were not coming from one's own contacts is a clear limitation in accounting for source influences on the perceived credibility of manipulative content shared on the platform. However, given that group chats in countries such as India and Brazil often consist of strangers united by a common interest (Bleize et al., 2021b; Kabha et al., 2019), *Join this Group* offers a unique insight into the efficacy of inoculating against misinformation when confronted with group norms and pressures (e.g., by positively reinforcing some choices and 'banning' players for other decisions).

This is consistent with research emphasising the influence of peer pressure in WhatsApp group chats (Brechwald & Prinstein, 2011) and its role in the formation of "WhatsApp vigilante" groupings (Banaji et al., 2019b). Promisingly, the present study is consistent with previous research (of which some is presented in this thesis), showing that inoculating interventions are effective in reducing the perceived reliability of false news (Roozenbeek & van der Linden, 2019) and boost people's confidence in their ability to spot and resist misinformation (Basol et al., 2020; Basol, Roozenbeek, et al., 2021), which, in turn, decreases self-reported willingness to share fake news items with their social networks (Roozenbeek & van der Linden, 2020). Though the intervention also affected the assessment of real news items, a pattern consistent with previous research (e.g., Guess et al., 2019), inoculated individuals' truth discernment still improved significantly more than those who were not inoculated. Naturally, the next step would be to inoculate towards psychological herd immunity. Though this has been hinted at through limited empirical studies, more research is required to examine whether and how attitudinal resistance can be best spread across individuals and whether it can not only keep up with the pace and depth at which misinformation travels. Once enough individuals are inoculated, misinformation would have a slim chance in going viral in the first place.

### Shifting towards the spread of inoculation

For instance, research suggests that attitude certainty is an essential determinant of attitudinal advocacy. That is, as people become more confident in their attitudes, they also become more willing to speak about them and even persuade others to adopt their views (Akhtar et al., 2013; Visser et al., 2003). On the other hand, some scholars suggest that inoculated individuals engage in vocalised counterarguing (post-inoculation talk; PIT) as a result of their heightened state of threat and vulnerability of having one's attitudes attacked (Compton & Pfau, 2009; Ivanov, Sellnow, et al., 2018). Furthermore, Compton and Pfau (2004) concluded that the negative relationship between confidence and threat suggested that inoculated individuals are more likely to turn to others for reassurance while *simultaneously* spreading inoculation-relevant information to their networks. Either way, it appears that "certainty is the catalyst that turns attitudes into action, bringing beliefs to life and imbuing them with meaning and consequence" (Tormala & Rucker, 2015). Consequently, while having established the role of attitude certainty in the process of resistance and its impact on the sharing of misinformation, research should extend these effects to the spread of inoculation as well.

While making interventions more accessible and scalable is a crucial factor in this process, I would argue that the prospect of making inoculation treatments *truly* generalisable should be the main focus of contemporary inoculation scholarship. Thus, two potential pathways towards psychological herd-immunity against misinformation emerge. First, developing an inoculation intervention that is scalable, effective cross-culturally, incorporates incentives to share it with others, and lastly, is effective against related yet untreated attitudes (a scarcely researched phenomenon regarded as "cross-protection"). In other words, rather than having to inoculate an individual against each topic and every argument, is it possible to push the boundaries of just how generalised the treatment is? The second avenue is one where the inoculated individual extends the benefits of sharing an inoculation message to the recipient. Whether it is driven by reassurance-seeking or advocative sharing, could people possibly be inoculated *vicariously*? Could the inoculated individual trigger a ripple effect of attitudinal resistance that can be passed along from one individual to another – vicariously and collectively inoculating towards societal resistance against misinformation while further boosting the 'spreader's' resistance? The next two chapters will explore these avenues, respectively.

### 3.7 Conclusion

A national representative data set was used to investigate the effectiveness of inoculation treatments in conferring attitudinal resistance against misinformation on online end-to-end encrypted private messaging platforms. The presented results are supporting the notion that gamified prebunking interventions are an effective and scalable means of reducing susceptibility to misinformation encountered within the context of WhatsApp. Furthermore, inoculated individuals also report being more confident in their assessments and less willing to share news items that employ manipulation strategies. This effect is maintained for at least one week after playing *Join this Group*. Additionally, this study demonstrates a mediating effect of confidence on reliability assessments and sharing intentions of fake items. In doing so, this research offers a substantial step toward a more nuanced understanding of the role of attitude certainty in the inoculation process and more importantly, how to build, strengthen, and spread attitudinal resistance against persuasion.

# 4 TOWARD PSYCHOLOGICAL HERD IMMUNITY: CROSS-CULTURAL EVIDENCE FOR TWO PREBUNKING INTERVENTIONS AGAINST COVID-19 MISINFORMATION

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

### 4.1 Abstract

Misinformation about the novel coronavirus (COVID-19) is a pressing societal challenge. Across two studies, one preregistered (n = 1,771 and n = 1,777), this chapter assess the efficacy of two "prebunking" interventions aimed at improving people's ability to spot manipulation techniques commonly used in COVID-19 misinformation, across three different cultural contexts (English, French, and German). The results suggest that Go Viral!, a novel 5-minute "prebunking" browser game, (a) increases the perceived manipulativeness of misinformation about COVID-19, (b) improves people's confidence in their ability to spot misinformation, and (c) reduces self-reported willingness to share misinformation with others. The first two effects remain significant for at least one week after gameplay. Consistent with the previous chapter, the findings also suggest a mediating effect of attitude certainty on sharing intentions of misinformation. We also find that reading real-world infographics from UNESCO improve people's ability and confidence in spotting COVID-19 misinformation (albeit with a smaller effect size than the game). Lastly, this chapter offers a variety of theoretical and practical contributions. By providing further support for the role of attitude certainty, comparing active and passive inoculation treatments, disentangling threat, and predicting resistance through motivational threat, this chapter offers a variety of theoretical and practical contributions to the inoculation scholarship. Limitations and implications are discussed.

### 4.2 Introduction

The previous two chapters took a closer look at some of the theoretical mechanisms underpinning the inoculation process and have provided a multitude of findings in the process. More specifically, this doctoral thesis thus far has established the efficacy of generalised and active inoculation interventions (*Bad News, Join this Group*) in conferring resistance against common misinformation techniques occurring on social media and private messaging apps. These interventions have incorporated differing socio-psychological factors such as political ideologies and group dynamics and have provided promising results for the efficacy of *actively* inoculating people with differing pre-existing attitudes
and degrees of previous exposure to misinformation. This is consistent with Compton and Ivanov (2013, p.276), who emphasised that "inoculation could be used as a single strategy to create resistance in individuals who hold the "right" attitude and at the same time persuade those with neutral and opposed positions to move in the "right" attitudinal direction". Furthermore, the findings give reason to argue that attitude certainty plays a pivotal role in generating and strengthening resistance. There is also some initial evidence that attitude certainty mediates sharing intentions of misinformation on WhatsApp, suggesting its potential use in the pursuit of psychological herd immunity. Yet, some fundamental questions remain unanswered.

This chapter aims to address the three-layered gaps in the current scientific understanding by shedding additional light on the study designs, theoretical foundations, and ultimately, the efficacy of the intervention against misinformation. Specifically, through the previous chapters examined the efficacy of gamified inoculation treatments in a controlled setting, they lacked a comparative intervention that would help identify whether active inoculation treatments do differ in their effectiveness from traditionally more passive interventions. Additionally, although Join this Group made use of a UK nationally representative data set, inoculation research across cultures is scarce and urgently needed. Thirdly, while the previous chapter examined the decay of the inoculation effect, it exposed participants to the same items assessed before and after treatment exposure. In other words, establishing whether generalised inoculation interventions extend resistance against previously unseen manipulation techniques is necessary. Specifically, within the context of the pandemic, testing these boundary conditions and establishing how to optimise the generation and spread of resistance is crucial. Which is why, lastly, this chapter aims to explore how an intervention designed to spread across arguments, topics, cultures, and between individuals may offer a pathway towards psychological herd immunity. To do so, clearer theoretical parallels must be drawn between the analogy and the interventions presented throughout this thesis. With a specific focus on threat this thesis will briefly revisit relevant theoretical foundations and gaps in the inoculation literature.

### *Revisiting the basics*

In recent years, the pre-emptive debunking ("prebunking") of misinformation has been put forward as a promising step towards building attitudinal resistance against misinformation. Prebunking is a key component of inoculation theory, often regarded as the "grandfather theory of persuasion" (Eagly and Chaiken, 1993: 561). Psychological inoculation is based on the biological analogy of an immunization process (McGuire, 1964). Similar to how exposure to a weakened dosage of a pathogen triggers the generation of protective antibodies, persuasion inoculation posits that a weakened persuasive argument will elicit motivation to equip oneself with protective arguments against it (McGuire and Papageorgis, 1961). Thus, both processes rely on the assumption that exposure to weakened forms of pathogens trigger an immunity-bolstering response. In cognitive inoculation, the immunization process often consists of the two theoretical elements, namely, a forewarning—which induces a

perceived threat to one's attitudes—and a pre-emptive refutation of the persuasive arguments (Compton, 2013). A large body of inoculation research has demonstrated the robustness and effectiveness of inoculation in conferring resistance across a multitude of topics (for a review, see Banas and Rains, 2010). A sense of threat is induced by warning participants about an imminent persuasive attack (e.g., future exposure to fake news), this heightened perception of vulnerability, in turn, is believed to motivate the generation of attitude-bolstering counterarguments against future persuasive attacks (Compton and Pfau, 2005).

## Essential theoretical building blocks

Thus, counterarguing and threat are regarded as the two key components of an inoculation treatment. McGuire's early theorising argued that "without such threatening stimulation, the person will be little motivated to assimilate the defensive material" (McGuire & Papageorgis, 1962, p.25). Additionally, some scholars have regarded threat as "the most distinguishing feature of inoculation" (Pfau, 1997, p.137) and have claimed that "inoculation is impossible without threat" (Compton & Pfau, 2005, pp.100-101). Furthermore, Compton and Ivanov (2012) noted that, "of the two components, threat is, arguably, the most important" (p.2). Therefore, one could suggest that the central role threat appears to play in the inoculation process is fundamentally tied to its function as a motivational catalyst toward resistance (Banas & Rains, 2017). Yet, and despite this assumption of centrality, the most recent metaanalysis of inoculation research found no significant relationship between perceived threat and actual attitudinal resistance (Banas & Rains, 2010). Considering that "threat is the sine qua non for inoculation-conferred resistance to influence" (Compton & Ivanov, 2012), this disparity between inoculation scholarship's commitment to the theoretical cornerstones and the empirical reality is perhaps one of the most striking and troubling findings of the meta-analysis. After all, McGuire himself never directly measured threat in any of his studies, leaving one of the theoretical pillars of inoculation to be mostly assumed (Compton, 2013; Banas & Richards, 2017).

Two immediate potential reasons may help disentangle this conundrum. First, it might be a matter of theoretical misconception. That is, perhaps the scholarship is mistaken in believing that threat is crucial to inoculation-induced resistance. Second, as first put forward by Banas and Richards (2017), it might be that this disparity in theoretical assumptions and empirical reality is driven by a flawed operationalisation of threat itself. Indeed, they suggest that the traditional measure of threat relies too heavily on the apprehension of threat rather than the motivation that is argued to drive the process of attitudinal resistance. To give an example, the authors argue that the words used in the traditional apprehensive threat measure (e.g., "scary", "dangerous", and "harmful") are more suited to reactions to physical or apprehension-inducing threat rather than the heightened understanding of one's attitudinal vulnerability. Indeed, scholars emphasise that threat must not be synonymous with fear and that it is possible for forewarned individuals to be defensive against attitudinal attack messages without being afraid or anxious about such an attack (Banas & Rains, 2017; Compton & Ivanov, 2014). However, despite generally good internal consistency demonstrated across a variety of studies,

perceived threat elicited by inoculation treatments was not significantly associated with resistance (Banas & Rains, 2010). Instead, it is possible for the traditional measure of threat to be significantly caused by the inoculation treatment without actually playing a part in conferring attitudinal resistance. Instead, alternative psychological states may mediate the effect of inoculation treatments on resistance (Banas & Rains, 2017). In turn, it is probable for inoculated individuals to be forewarned and on guard about an imminent attack message without being afraid or anxious. And while this notion that the heightened sense of attitudinal vulnerability is not dependent on fear is recognised (Compton & Ivanov, 2014; Pfau, 1995), the continued use of a scale that relies on synonyms for fear is widespread.

# Why apprehension is not the solution

Scholars demonstrated the independent effects of fear and cognition on persuasive outcomes (Dillard, 1994). Though the theoretical assumptions in inoculation research posit that threat makes attitude vulnerability salient and motivates an active process of counterarguing, fear and subsequent mechanisms of attitudinal protection appear to be rooted in withdrawal, escape, and avoidance instead (Nabi, 2002; Shen & Dillard, 2007). Similarly, it is commonly argued that the response of fear to a persuasive argument is primarily linked to behaviour-inhibiting systems instead of behaviouractivating systems (Dillard & Anderson, 2004). Thus, it is possible that traditionally apprehensive inoculation messages trigger inhibiting processes that result in less resistance. However, it is important to note that while fear in itself may not be beneficial for generating attitudinal resistance, literature exploring the role of threat-induced anger suggests a potentially powerful impact of outrage on resistance. In other words, whether it is isolated anger or in form of a fear-based reaction, anger may play an effective motivational role in the heightened desire to protect ones attitudes (e.g. Ivanov et al., 2009). This further reinforces the frequently- mentioned yet scarcely addressed need for understanding the role of affect in inoculation. Under which circumstances anger-induced, heightened states of motivation lead to effective resistance rather than avoidant cognition remains unclear. Overall, Compton and Ivanov (2014) conclude that researchers "have not taken many steps towards determining what threat is" (p.18).

#### Bringing motivation back into threat

Following Compton and Ivanov (2012), who stated that threat "should motivate the process(es)...that ultimately leads to resistance" (p.2), Banas and Richard's (2017) suggested creating an alternative account of threat by disentangling motivation and apprehension. This re-conceptualisation of threat as motivation to defend one's beliefs appears to be theoretically more consistent with inoculation theory. Compton and Pfau (2005) defined threat as the "catalyst to change" necessary to "motivate the work needed to strengthen an attitude" (p.100).

If threat was to be approached as a challenge rather than a source of danger and fear, it would invoke the necessary motivation to defend one's attitudes instead of leading to anxiety, intimidation, or fear. This is consistent with scholars arguing that threat "unleashes the counterarguing process" (Pfau, et al., 2010; p.3) and motivates the inoculated individual to engage in the counterarguing process of raising and refuting attitude-specific arguments (Pfau et al., 2005). In short, the rhetoric around the role of threat presumes that it motivates counterarguing (Compton & Ivanov, 2012). In fact, Banas and Richard (2017) find that operationalising threat as motivation (hereafter termed motivational threat) is conceptually different from the traditional threat measure, significantly more predictive of attitudinal resistance, and lastly, less related to fear. The authors conclude by emphasising that the motivational threat measure is a more suitable approach to understanding the mechanisms underlying the affective component of the inoculation process. As previously argued, this is in line with the tendency of inoculation research to rely on an overtly cognitive approach to resistance to persuasion.

# Attitudinal inoculation during a global health crisis

While highlighting the apparent discrepancy between theoretical assumptions and empirical reality in threat is important, it remains unclear whether and how socio-cultural factors and current affairs might influence the way in which threat is perceived. For instance, Banas and Richard (2017) found the above-mentioned findings when applied to the context of a *9/11 Truth* conspiracy. One could argue that other topics or contexts might be perceived as more relevant, urgent, and indeed, fearful to the inoculated individual. The SARS-CoV-2 (COVID-19) pandemic offers a good opportunity to compare the concepts of threat (motivational, apprehensive) as well as to test the efficacy of a generalised and therapeutic inoculation intervention in a highly contested and uncertain context. However, such an examination of inoculation "in the wild", or in other words, in the midst of a global health crisis, comes with a set of challenges mostly unaccounted for in inoculation research thus far.

For instance, much like the virus itself, misinformation about the disease has spread widely on social media, ranging from fake "remedies and cures", such as eating garlic or injecting bleach to elaborate conspiracy theories behind the cause of COVID-19 (BBC News, 2020). The World Health Organization (WHO) has warned of an "infodemic" (Zarocostas, 2020), and some have urged that the prevalence of misleading information around the virus might be "the most contagious thing about it" (Kucharski, 2020). Misinformation about COVID-19 affects people's willingness to comply with evidence-based health regulations (Imhoff and Lamberty, 2020) and has been associated with lower vaccine uptake intentions around the world (Roozenbeek, Schneider, *et al.*, 2020). Thus, a traditionally passive inoculation treatment (i.e., text-based pre-emptive refutations) cannot adequately keep up with the rapid "fake news engine" of continuously changing content and arguments around COVID-19 conspiracies. Although the previous two chapters have provided support for the efficacy of generalised inoculation interventions against manipulation techniques (Basol et al., 2020), little is currently known about whether active forms of inoculation treatments are indeed superior to passive forms (Banas and Rains, 2010). Especially with the case for attitude certainty growing, more research is needed to compare their respective abilities to generate confidence and prevent inoculation effects from decaying over time. This is important because if people are not sufficiently confident in their discernment attitudes they may be easily persuaded (Basol, Roozenbeek and van der Linden, 2020). Considering research that suggests that the more certain individuals are of their attitudes, the more likely it is to guide behaviour (Rucker and Petty, 2004), resist persuasion (Tormala and Petty, 2002), and persist over time (Bassili, 1996; Tormala, 2016), it remains unknown whether these effects hold true in generalised and therapeutic interventions applied to the unprecedented and uncertain context of a global pandemic. In other words, are therapeutic inoculation treatments effective in a politicised global health crisis characterised by constant changes in available information and scientific understanding, exposure to misinformation and conspiracy theories, and public opinion on the issue?

To add to the problem, information about COVID-19 is not easily classified as either true or false. Especially given the continuously developing scientific understanding of the virus (Vraga and Bode, 2020), information around it can range between various degrees of unverified or disproven, making a clear identification of what counts as "fake news" difficult. Nevertheless, recent surveys suggest that misinformation about COVID-19 is prevalent. To give an example, results by Pew reported that almost half of the sample (48%) had been exposed to misleading and false information about the coronavirus in the United States (Schaeffer, 2020). Moreover, among those who reported exposure to misinformation, the majority claimed to see such information on a daily basis. Frequent exposure to misinformation is particularly dangerous as repetition increases reliance on the false information (Fazio et al., 2015; Pennycook, Cannon and Rand, 2018; Effron and Raj, 2019). As the supply of evidence-based interventions remains low (Pennycook et al., 2020), it is critical to explore how the spread of misinformation around COVID-19 may be mitigated. In the absence of a vaccine, the mitigation of this crisis relies in large part on non-pharmaceutical interventions that leverage insights from the social and behavioural sciences (Depoux et al., 2020; Van Bavel et al., 2020; van der Linden, Roozenbeek and Compton, 2020). Thus, honing in on the theoretical opaqueness and teasing out the mechanisms to facilitate conferring, strengthening, and attitudinal resistance is crucial to mitigating this crisis.

### THE PRESENT RESEARCH

Across two studies, this chapter aims to test the effectiveness of two prebunking interventions designed at improving people's ability to spot misinformation about COVID-19. While doing so, this research also aimed to further tease out clearer roles within gamified inoculation that correspond with the analogical foundations of inoculation theory. Thus, while aiming to compare two treatments, replicate the findings from the previous two chapters, and test the efficacy of inoculation cross-culturally, this research also introduces an updated operationalisation of threat and examines the decay

effect and boundaries of *generalised* inoculation with previously unseen items. By incorporating the recent findings of the role of attitude certainty in the inoculation process, this research aims to take a critical look at the theoretical foundations and current practical executions of inoculation to further establish how resistance to misinformation can be strengthened, and more importantly, shared. Disentangling and identifying the motivations behind inoculated individuals sharing behaviours, and more specifically, the passing-on of the inoculation treatment itself will constitute one pathway towards establishing societal immunity against misinformation. Particularly within the context of the pandemic, the mitigation of this crisis relies in large part on non-pharmaceutical interventions (e.g., adherence to health guidelines, vaccine *uptake*) and thus, must leverage insights from the social and behavioural sciences (Depoux *et al.*, 2020; Van Bavel *et al.*, 2020; van der Linden, Roozenbeek and Compton, 2020). Once again, inoculation theory has the potential to effectively apply itself to a societally pressing issue. Thus, this chapter represents the joined effort with the UK Cabinet Office and the WHO to develop, test, and launch pre-emptively debunking interventions against COVID-19 misinformation.

The first intervention is Go Viral!, a novel and freely available 5-minute choice-based browser game similar in design to other "fake news" games such as Bad News and Join this Group (Roozenbeek and van der Linden, 2020; Harijani, Basol et al., in press). At the Social Decision-Making Research Lab, we created Go Viral! (www.goviralgame.com) in collaboration with the UK Cabinet Office and WHO<sup>12</sup>, media platform DROG, and design agency Gusmanson to expose three manipulation techniques commonly used in COVID-19 misinformation: fearmongering, using fake experts, and spreading conspiracy theories (WHO, 2020; Zarocostas, 2020). Go Viral!, was part of WHO's "Stop the Spread" campaign against COVID-19 misinformation, is available in three languages (English, French, German) and, at the time of the write-up, has been played more than 1,4 million times since its launch in early October 2020. Similar to Bad News, Go Viral! functions as an active inoculation against future manipulation attempts by pre-emptively warning and exposing people to weakened doses of misinformation and letting them generate their own "mental antibodies" (van der Linden and Roozenbeek, 2020). In the game, at the very beginning, players are explicitly warned (threat) about the ill-intent behind manipulative misinformation coupled with a call to learn about the techniques used by manipulators (i.e., pre-emptive refutation). Next, players start out by browsing their (fictitious) social media feed and are slowly lured into an echo chamber where misinformation and outrageevoking content about COVID-19 are common (these scenarios are aimed at eliciting threat and motivation). This choice and development of the issues raised in the fictitious content was informed by insights provided by the collaborators. Across three scenarios, players are exposed to weakened doses of three common manipulation techniques.



Figure 14: Screenshot of in-game threat element used at the beginning of 'Go Viral!'.

In the first scenario, "The Fearmongerer", players create a social media post by using moral-emotional language and watch it go viral. The use of moral-emotional language is known to enhance the virality of social media content (Brady et al., 2017; Acerbi, 2019; Berriche and Altay, 2020). After this initial success, they are asked to join Not Co-Fraid, a group of online "truth tellers". In the second scenario, "My Imaginary Expert", players start sharing content in the group as Not Co-Fraid's newest member, but soon find out that their credibility remains low. They are then prompted to look for make-believe experts who will back up their false claims, such as Dr Hyde T. Paine from the "University of Life". By giving Not Co-Fraid group members the illusion that their content is backed up by experts, players gain even more likes, and are eventually asked to become a Not Co-Fraid moderator. This scenario relies on impersonation and the fake expert technique (Cook, Lewandowsky and Ecker, 2017; Roozenbeek and van der Linden, 2019). In the final scenario, "Master of Puppets", players are encouraged to go even further. They create their own COVID-19 conspiracy theory by picking a target (e.g., a large NGO, the government, or one Bob from New York), accusing it of shady practices, and connecting the dots. This conspiracy theory is so successful that nationwide protests erupt against the player's target. Conspiracy theories have featured heavily around COVID-19 (van der Linden, Roozenbeek and Compton, 2020) and have been linked to violent intentions (Jolley and Paterson, 2020) and reduced willingness to comply with public health measures (Roozenbeek, Schneider, et al., 2020). The game ends by showing players the negative consequences of their actions, for example a friend that players met early in the game telling them that their friendship is over after the player's actions. Figure 15 shows the Go Viral! landing page and game environment.



Figure 15: Go Viral! landing page (left) and game environment (middle and right).

The second real-world intervention consists of a series of infographics about COVID-19 misinformation developed by UNESCO (UNESCO, 2020). As part of its #ThinkBeforeSharing prebunking campaign, UNESCO created a series of images that can be easily shared on social media and that explain how misinformation about COVID-19 is created and spread. These infographics were designed with input from inoculation researchers<sup>13</sup>. Figure 16 shows several examples.



Figure 16: UNESCO infographics.

This study leverages the opportunity of the public availability of both interventions to test several key hypotheses pertaining to the effectiveness of real-world "prebunking" as a way to reduce susceptibility

<sup>&</sup>lt;sup>13</sup> Upon our request, UNESCO made all infographics available in English, French, and German so we could adequately compare both interventions cross-culturally.

to COVID-19 misinformation before any vaccines were available. To date, no published research has tested different types of inoculation and prebunking interventions alongside each other. Crucially, it has been hypothesized that so-called "active" inoculation (e.g., in the form of a game) is more effective at conferring attitudinal resistance than "passive" inoculation (i.e., through reading, see McGuire and Papageorgis, 1961; Banas and Rains, 2010; Roozenbeek and van der Linden, 2018). This study is the first to assess the effectiveness of passive and active interventions alongside each other. In addition, these two studies presented in this chapter are the first to test the effectiveness of prebunking interventions in the context of COVID-19 misinformation. Following recommendations to maximise the generalisability of intervention effects (O'Keefe, 2015), this study also make use of the fact that both *Go Viral!* and the UNESCO infographics are available in English, French, and German to test the effectiveness of each intervention across different.

Finally, given the decay of persuasion effects (Maertens et al., 2020), the long-term effectiveness of both inoculation interventions was evaluated in a follow-up one week after the initial study. Importantly, and contrary to previous research so far, the follow-up presents new and previously unseen items, allowing to examine the boundaries of the "blanket of protection" conferred through generalised inoculation. Moreover, by incorporating different forms of threat (apprehensive, motivational) alongside attitude certainty and sharing intentions, this study is the first to examine the role of both threat and confidence in the process of strengthening and sharing therapeutic inoculation. Doing so will significantly contribute to the ongoing efforts to establish clear and feasible directions towards psychological herd immunity (Lewandowsky & van der Linden, 2021).

This chapter addressed the above-mentioned questions across two high-powered and large-sample studies. All the necessary information needed to replicate our findings and methods, including our datasets, Qualtrics surveys, the full list of items (social media posts), preregistrations, supplementary tables, figures and analyses, and the analysis and visualisation scripts can be found on the Open Science Framework (OSF) page: <u>https://osf.io/mbqwj/</u>. Both studies were approved by the Cambridge Psychology Research Ethics Committee (PRE.2020.035).

### 4.3 Go Viral! - Study 1

Study 1 used a voluntary in-game participation, aiming to establish the efficacy of *Go Viral!* in conferring resistance against fake news items presented before and after gameplay. Thus, this stage focused on perceived manipulativeness of news items (3 real, 3 fake) and tested whether the inoculation effect differed across demographics.

## 4.4 Methods

## Procedure

For our first study, a voluntary pre-post survey within the *Go Viral!* game, following the within-subject paradigm developed by Roozenbeek and van der Linden (2019) was implemented. At the start of the game, players were asked to participate in a scientific study. If they consented, they were shown three fake and three real social media posts (in the form of Tweets) relating to COVID-19, and asked to rate the manipulativeness of each post on a 1-7 Likert scale (following Saleh *et al.*, 2020). After completing the game, players were asked to participate in the second part of the study. Upon agreeing to do so, they were again asked to rate the manipulativeness of the same social media posts that they saw in the pre-test and to answer a series of demographic questions (gender, age group, education level, geographic region and political ideology). Participants were not compensated financially for their participation.

The three fake social media posts each make use of a manipulation technique learned in the game (the use of moral-emotional language, using fake experts, and conspiratorial reasoning), and were taken from fact-checking websites such as FullFact and the WHO's COVID-19 Mythbusters page (World Health Organization, 2020). The three real posts are Tweets about COVID-19, taken from the Twitter accounts of reputable news sources (BBC News, AP, Reuters)<sup>14</sup>. To avoid potential source confounds, all source information for both the real and fake posts was blacked out so that participants could not see the source of the information but could only assess them by looking at the wording and language use. Figure 17 shows what the survey looks like within the game environment.

<sup>&</sup>lt;sup>14</sup> All items used in this study can be found in the "items" folder on the OSF: <u>https://osf.io/mbqwj/files/</u>



Figure 17: In-game survey screenshots: start of the survey (left), consent form (middle) and a social media post (right).

## Sample

Between 27 October and 26 November 2020, a total of 2,634 complete pre-post survey responses were collected within the *Go Viral!* game environment. 863 participants indicated being under 18 years old and were excluded as per the ethics approval, leaving a total sample of N = 1,771.52.9% of the sample identified as male (43.0% female, 1.8% other, 2.3% prefer not to say). 53.6% indicated being between 18 and 34 years of age, and 36.3% reported having a university bachelor's degree. The sample also skewed politically left (M = 3.07, SD = 1.24). In terms of geographic region, most study participants were from Europe (59.3%) and North America (22.7%). See online Supplementary Table S1 for the full sample composition.

This study design allows to test the following hypotheses pertaining to the effectiveness of *Go Viral!* as a way to improve people's ability to spot misinformation about COVID-19:

 $H_1$ : People who play *Go Viral!* will rate misinformation about COVID-19 as significantly more manipulative after gameplay.

**H**<sub>2</sub>: People who play *Go Viral!* will be able to distinguish real news and misinformation about COVID-19 more accurately after gameplay.

In addition, the following null hypothesis was tested:

 $H_{0,realnews}$ : People who play *Go Viral!* will not rate real news as significantly more manipulative after gameplay.

#### 4.5 Results

To test hypothesis **H**<sub>1</sub>, a paired-samples *t*-test on the *averaged* pre- and post-manipulativeness scores for the three misinformation items was conducted <sup>15</sup>. The results suggest that participants rate misinformation about COVID-19 as significantly more manipulative after playing *Go Viral!*  $(M_{fakenews,pre} = 5.61, M_{fakenews,post} = 6.07, M_{diff} = 0.46, 95\%$  CI [0.419, 0.502], t(2,1770) = 21.88, p <0.001, d = 0.52, 95% CI [0.470, 0.569]). The same results are found for the individual emotion item  $(M_{emotion,pre} = 5.40, M_{emotion,post} = 5.83, M_{diff} = 0.43, 95\%$  CI [0.367, 0.495], t(2,1770) = 13.21, p < 0.001,d = 0.31, 95% CI [0.266, 0.362]), the fake expert item ( $M_{fakeexpert,pre} = 5.45, M_{fakeexpert,post} = 6.15, M_{diff}$ = 0.70, 95% CI [0.634, 0.763], t(2,1770) = 21.11, p < 0.001, d = 0.50, 95% CI [0.452, 0.551]), and the conspiracy item ( $M_{conspiracy,pre} = 5.98, M_{conspiracy,post} = 6.23, M_{diff} = 0.25, 95\%$  CI [0.196, 0.308], t(2,1770) = 8.77, p < 0.001, d = 0.21, 95% CI [0.161, 0.255]). These results support hypothesis **H**<sub>1</sub>.

To test hypothesis  $H_2$ , a paired samples *t*-test was conducted on the pre- and post-gameplay difference for the difference in manipulativeness scores between fake news and real news (i.e., the level of "veracity discernment", or fake news manipulativeness minus real news manipulativeness), before and after playing. Doing so gives a significant post-gameplay increase in veracity discernment, showing that *Go Viral!* players are better able to distinguish real and false information about COVID-19 after playing, in support of hypothesis  $H_2$  ( $M_{discernment,pre} = 2.98$ ,  $M_{discernment,post} = 3.41$ ,  $M_{diff} = 0.43$ , 95% CI [0,374, 0.487], t(2,1770) = 14.94, p < 0.001, d = 0.36, 95% CI [0,307, 0.403]).

For real news, no significant difference in overall pre-post manipulativeness scores is found  $(M_{realnews,pre} = 2.63, M_{realnews,post} = 2.66, M_{diff} = 0.03, 95 \%$  CI [-0.0180, 0.0782], t(2,1770) = 1.23, p = 0.22, d = 0.03, 95 % CI [-0.0174, 0.0758]). Furthermore, no significant pre-post differences for two out of three real news items are found either (both ps > 0.13), and a small but significant increase in the perceived manipulativeness of one real item ( $M_{asia,pre} = 2.83, M_{asia,post} = 2.91, M_{diff} = 0.09, 95 \%$  CI [0.0192, 0.157], t(2,1770) = 2.51, p = 0.012, d = 0.06, 95 % CI [0.013, 0.106]). A Bayesian paired samples *t*-test for the averaged real news items gives a Bayes factor of  $BF_{10} = 0.057$  (error % = 0.057).

<sup>&</sup>lt;sup>15</sup> Due to the low number of items, the averaged indices for the fake and real news items have relatively low internal consistency ( $a_{fake} = 0.53$  and  $a_{real} = 0.71$ ). We therefore also report item-level alongside averaged results.

 $(0.046)^{16}$ , indicating strong support for the null hypothesis **H**<sub>0,realnews</sub> (see van Doorn et al., 2020). Figure 18 shows the results for fake news and real news in a bar graph.



Figure 18: Bar graph of the perceived manipulativeness of fake news (left) and real news (right), averaged and per individual item. Error bars show 95% confidence intervals.

As a robustness check, Bayesian paired samples *t*-tests for the averaged manipulativeness scores for veracity discernment, real news and fake news, as well as for each individual item was conducted. Doing so shows strong support under the Bayesian framework for both hypotheses  $H_1$  and  $H_2$ , as well as for the null hypothesis  $H_{0,realnews}$ ; see Supplementary Table S2 for a full overview. Finally, to check for interaction effects, a linear regression was conducted with the difference in pre-post veracity discernment as the dependent variable, and gender, age group, education level, political ideology and being from Europe (as this was the largest single geographic region of origin in our sample) as

<sup>&</sup>lt;sup>16</sup> We used two different Cauchy priors for the Bayesian *t*-tests reported here, centred around 0, with two width parameters: 0.707, (the default width parameter recommended by van Doom et al., 2020; corresponding to an ~80% that the effect size lies between -2 and 2), and 0.16, which corresponds to a 79% chance that the effect size lies between -0.50 and 0.50 (in line with previous effect sizes reported in similar pre-post research designs, see Roozenbeek and van der Linden, 2019). The results reported here are for the width parameter of 0.707. With a width parameter of 0.16, a Bayesian paired samples *t*-test gives a Bayes factor of  $BF_{10} = 0.239$  (error % = 0.019), also indicating strong support for the null hypothesis **H**<sub>0,realnews</sub>.

covariates. No significant effects were found (all ps > 0.082), except for political ideology (p = 0.006), in the sense that people who identify as left-wing display a higher post-pre inoculation effect in terms of veracity discernment than people who identify as right-wing. See OSF page, Supplementary Table S3.

# 4.6 Discussion – Study 1

In a large-sample in-game survey experiment, Study 1 showed that people who play Go Viral!, irrespective of their demographic background (aside from political ideology), found misinformation about COVID-19 significantly more manipulative after playing than before, whereas their assessment of real news about the virus did not change in a meaningful sense. The effect sizes are in line with previous studies that used a similar design (Roozenbeek and van der Linden, 2019), and are particularly encouraging considering these are within-subjects effects. Thus, it can be argued that this suggests that inoculation treatments are not constrained to issue topics on which individuals have preexisting attitudes but rather, that inoculating against real-time crises, where the societal and scientific narrative is changing and evolving. Moreover, these results also provide preliminary support for the effectiveness of active, generalised, and prophylactic inoculation treatments on topics that are perceived as uncertain, anxiety-inducing, or fearful. Regarding the existing literature on the role of threat in the inoculation process (Banas & Rains, 2010; Compton, 2013l; Banas & Richards, 2017), the following question then concerns whether the threat elicited in Go Viral! predicts resistance to COVID-19 misinformation. Although this study allowed us to leverage the popularity of the Go Viral! game to collect survey responses "in the wild", it suffers from selection bias, and it was not possible to test the game against other interventions aimed at reducing susceptibility to misinformation about COVID-19. In addition, to avoid overburdening study participants within the game, only a total of six items and one outcome measure were included. These issues are addressed in Study 2.

# STUDY 2

### 4.7 Methods

For this study, a pre-registered randomised controlled trial on *Prolific Academic* was conducted with three conditions (an active condition, a "passive" Infographic condition, and a neutral control condition), and in three languages: English (using a national sample of the United Kingdom), French,

and German<sup>17</sup>. Additionally, this study aimed to further examine the role of attitude certainty and motivational threat in the generation, strengthening, and spreading of resistance against misinformation. The inoculation condition involved playing the Go Viral! game. The Infographic condition involved reading through the UNESCO infographics<sup>18</sup>. The control condition involved playing Tetris for approximately the same amount of time it takes to complete the Go Viral! game. Tetris was chosen as a control condition for several reasons: 1) it has been used as a control condition in previous studies on inoculation games (Basol, Roozenbeek and van der Linden, 2020; Maertens et al., 2020; Roozenbeek and van der Linden, 2020); 2) it is in the public domain; and 3) it is a simple game with a flat learning curve. At the start of the study, participants performed an item rating task, in which they were randomly shown nine real and nine fake social media posts (in the form of tweets, in English, French or German) about COVID-19. Six of these 18 items were the same as those used in Study 1, the other 12 were selected using the same procedure as described in Study 1<sup>19</sup>. In total, participants thus saw nine real news posts (not containing misinformation) and nine misinformation posts (three for each manipulation technique: fearmongering, fake experts, and conspiracy). As in Study 1, all source information was blacked out to avoid source confounds (Maertens et al., 2020). For each post, participants were asked to rate the following statements on a 1-7 scale (1 being "strongly disagree" and 7 being "strongly agree): 1) this post is manipulative; 2) I am confident in my assessment of this post's manipulativeness; 3) I would share this post with people in my network. Figure 19 shows what this looks like within the survey environment.

<sup>&</sup>lt;sup>17</sup> The pre-registrations can be found here: <u>https://aspredicted.org/28sr5.pdf</u> (UK), <u>https://aspredicted.org/blind.php?x=aa8mz2</u> (German), <u>https://aspredicted.org/blind.php?x=nj3iw7</u> (French). Aside from the sample size (discussed in the "Sample" section), we report no significant deviations from the pre-registration.

<sup>&</sup>lt;sup>18</sup> We deliberately did not set a minimum time for participants to read the infographics. Our reason for doing so is that we wanted the intervention to simulate a real social media environment as much as possible, in the sense that Infographics condition participants had to scroll through the page to read the infographics, much like one may come across them on one's Twitter feed or Facebook timeline.

<sup>&</sup>lt;sup>19</sup> All items (for the UK, French and German studies, the UK follow-up and the pilot study) can be found in the "items" folder on the OSF: <u>https://osf.io/mbqwj/files/</u>

| Cid yru koar that the<br>tyte in the an mRNA va<br>teell into your system | UK reports 4,044 COVID-19 cases on Monday compared<br>with 5,693 on Sunday |            |   |         |   |                |   |   |          |          |          |         |        |      |           |
|---|--|------------|---|---------|---|----------------|---|---|----------|----------|----------|---------|--------|------|-----------|
|   | Strongly   | fisagree . |   | Neutral |   | Strongly agree |   |   |          |          |          |         |        |      |           |
|   | 1  | 2          | 3 | 4       | 5 | 6              | 7 |   | Strongly | disagree |          | Neutral |        | Stor | gly agree |
| This post is<br>manipulative  | 0  | 0          | 0 | 0       | 0 | 0              | 0 | This post is manipulative   | 1        | 2        | 3 I<br>O | 4       | 5<br>O | 6    | 7         |
| I am confident in my<br>judgment about this<br>post's<br>manipulativeness | 0  | 0          | 0 | 0       | 0 | 0              | 0 | I am confident in my<br>judgment about this<br>post's<br>manipulativeness | 0        | 0        | 0        | 0       | 0      | 0    | 0         |
| I would share this post<br>with people in my<br>network                   | 0  | 0          | 0 | 0       | 0 | 0              | 0 | I would share this post<br>with people in my<br>network                   | 0        | 0        | Ó        | 0       | 0      | 0    | 0         |

Figure 19: Examples of a manipulative (left) and real (right) social media post from the item rating task (Study 2).

After completing this item rating task, participants were randomly assigned to one of the treatment conditions (active inoculation or Infographics) or the control condition. Participants in the Infographics condition were given one manipulation check after reading the infographics (i.e., whether the infographics they saw contained a certain hashtag). Participants in the inoculation condition were given two manipulation checks: they had to provide a password that they were given at the end of the *Go Viral!* game, and they were asked what happened at the end of the final scenario. After the manipulation check, participants were given two tasks in a random order<sup>20</sup>: 1) a set of questions about perceived motivational threat adjusted to the context of misinformation about COVID-19 (Miller *et al.*, 2013; Richards and Banas, 2018)<sup>21</sup>; and 2) the same item rating task that participants were given in the pre-test (i.e., the post-test). After these two tasks, participants answered a series of questions: the "vigilance" measure from the Reuters Digital News Report (a measure of the extent to which people are concerned about the accuracy and source reliability of the news that they consume, with responses ranging from "never" to "frequently" on a 4-point scale; see Newman *et al.*, 2020); self-perceived resistance against misinformation (1-7; see Ivanov *et al.*, 2017); people's willingness to share

<sup>&</sup>lt;sup>20</sup> The presentation of threat items was counterbalanced to avoid order-effects on the post-rating task.

<sup>&</sup>lt;sup>21</sup> A confirmatory factor analysis was conducted for Richards and Banas' (2018) motivational threat measure. Following their use of Hu and Bentler's (1990) two-index criteria of comparative fit index (CFI > .95) and the standardized root mean square residual (SRMR < .08), the model demonstrated good fit,  $\chi^2$  (34) = 9.48, p < .01, CFI = .94, SRMR = .048. Thus, the items exhibited unidimensionality.

the game (*Tetris/Go Viral*!) or the UNESCO infographics on social media accounts (1-7) and in real life (1-7); whether people had ever been infected with COVID-19 (yes/no/unsure/prefer not to say; see Dryhurst *et al.*, 2020); how worried they are about COVID-19 (1-7; see Dryhurst *et al.*, 2020); and whether they would get vaccinated against COVID-19 if a vaccine became available (yes/no; see Roozenbeek, Schneider, *et al.*, 2020).

And finally, participants were asked several standard demographic questions: birth year, gender, education, and political ideology (1 being "very left-wing" and 7 being "very right-wing"). For the UK sample only<sup>22</sup>, participants were recontacted one week after their participation in the study for a brief follow-up, in which they were asked to perform the same item rating task (with manipulativeness, confidence, and willingness to share as outcome measures) for twelve *new* and previously *unseen* social media posts (6 real and 6 fake, or two fake posts per misinformation technique learned in the *Go Viral!* game).

With the above in mind, the pre-registered hypotheses for Study 2 were the following:

 $H_3$ : Participants in both treatment conditions (*Go Viral*! and the UNESCO infographics) will assess the manipulativeness of real and false information more accurately than the control condition, in all 3 languages tested.

 $H_4$ : Participants in both treatment conditions will be more confident in their manipulativeness assessments than the control condition, in all 3 languages tested.

**H**<sub>5</sub>: Participants in both treatment conditions will be less willing to share false information with others in their network than the control condition, in all 3 languages tested.

 $H_6$ : Participants in the active inoculation condition (*Go Viral!*) will be more willing to share the treatment [with others] than the Infographics condition, in all 3 languages tested.

 $H_{7a}$ : One week after exposure to the intervention, participants in the inoculation conditions will display minimal decay of the inoculation effect (for manipulativeness, confidence & sharing), in all 3 countries tested.

 $H_{7b}$ : One week after exposure to the intervention, participants in the infographics condition will display significant decay of the inoculation effect (for manipulativeness, confidence & sharing).

<sup>&</sup>lt;sup>22</sup> Due to budgetary constraints, we were unable to do a follow-up for all three languages. The UK results were of primary interest given that the intervention was launched with the UK government.

Finally, based on the pre-registered exploratory analyses on motivational and traditional threat, it can be hypothesised that,  $H_8$ : Perceived threat about COVID-19 misinformation is significantly higher in the inoculation condition compared to the Infographics and control conditions. The flowchart in Figure 20 shows the study's design schematically.



Figure 20: Study flowchart.

## Sample

Participants were recruited via Prolific Academic (Peer et al., 2017). First, a pilot study was conducted (n = 231) as a pre-test, in order to validate our item sets<sup>23</sup>. Next, the full study was run in three different languages: one with a national sample of the UK (in English), one in French, and one in German<sup>24</sup>. Participants were paid GBP 1.75 for their participation. UK participants were invited to participate in a short follow-up one week after the initial study, for which they were paid an additional GBP 0.25. Participants in the Infographics and Go Viral! conditions were subjected to one (for Infographics) or two (for Go Viral!) attention checks. As per our pre-registration, participants who failed one (in the Infographics condition) or both (in the Go Viral! condition) manipulation checks were excluded from the analysis. The pre-registrations for the French and German studies also stated using Facebook and/or Twitter as an a priori inclusion criteria. A priori power analysis using G\*Power with an effect size of d = 0.40, 95% power, 3 groups and 3 measurements (pre - post - follow-up) gives a desired sample size of n = 261 per language to detect a main effect. However, because the effect size is expected to be smaller for the confidence and sharing measures (see Roozenbeek and van der Linden, 2020), and in line with our pre-registration, this study aimed to recruit 900 participants for each language. Unfortunately, due to problems with survey completion (an unexpected number of participants did not provide the correct game completion password), the initial target sample size was not met. Instead, the final sample consists of n = 710 valid participants for the UK study, n = 610 for the French study, and n = 457 for the German study, for a total of N = 1,777. In total, 606 out of 701 UK respondents participated in the one-week follow-up study (86% retention). See Supplementary Table S4 for the full sample composition by each country.

<sup>&</sup>lt;sup>23</sup> The full dataset for the pilot study, in which we pre-tested a set of 15 fake and 15 real items, can be found on the OSF. In total, 18 items (9 real and 9 fake) were used in the pilot study; 12 were rejected. For the follow-up study, we used 6 real and 6 fake items not tested in the pilot study.

<sup>&</sup>lt;sup>24</sup> For the German and French studies, we used "German-speaking" and "French-speaking" as inclusion criteria; pre-selecting for German or French nationality did not yield a large enough pool of Prolific participants.

#### 4.8 Results

In the interest of clarity and brevity, the results will primarily be reported for the pooled sample (UK, French, and German put together)<sup>25</sup>. The results per country are all directionally similar, but differences in significance levels for each country are reported where appropriate. All item-level statistics for each outcome variable of interest (descriptive statistics and ANOVA results), pooled and by country, can be found in Appendix.

### Manipulativeness

For the pooled sample, a one-way between-subjects ANOVA shows a significant effect of condition (control, *Go Viral!*, Infographics) on the pre-post intervention difference in the perceived manipulativeness of fake news about COVID-19 (F(2,1774) = 51.69, p < 0.001,  $\eta^2 = 0.055$ , d = 0.48)<sup>26</sup>. A Tukey HSD post-hoc comparison shows that the pre-post difference in perceived manipulativeness for the *Go Viral!* condition was significantly higher than the control condition (M = 0.45 vs M = 0.08,  $M_{diff} = 0.37$ , 95% CI [0.28, 0.46],  $p_{tukey} < 0.001$ , d = 0.56) and the Infographics condition (M = 0.45 vs M = 0.08,  $M_{diff} = 0.18$ ,  $M_{diff} = 0.27$ , 95% CI [0.18, 0.36],  $p_{tukey} < 0.001$ , d = 0.41). The same effect is found for the infographics condition compared to the control condition (M = 0.18 vs M = 0.08,  $M_{diff} = 0.10$ , 95% CI [0.02, 0.19],  $p_{tukey} = 0.015$ , d = 0.17), indicating that both playing the *Go Viral!* game and reading through the UNESCO infographics significantly increase the perceived manipulativeness of COVID-19 fake news. These results are similar (and significant) in all 3 countries; see Appendix for a full overview as well as item-level statistics. Figure 21 shows these results in a violin plot.

<sup>&</sup>lt;sup>25</sup> See Supplementary Tables S24-S28 for invariance testing between countries for both the real and fake items for the manipulativeness, confidence, and sharing measures.

<sup>&</sup>lt;sup>26</sup> A reliability analysis gives acceptable internal consistency for the manipulativeness measure for the fake items (M = 5.26, SD = 0.94, Cronbach's  $\alpha = 0.77$ ) and the real items (M = 3.13, SD = 1.02, Cronbach's  $\alpha = 0.81$ ), so we report the results for the average of both real and fake items. See also Supplementary Table S7.



Figure 21: Violin plot with jitter of post-pre manipulativeness scores of fake news posts (all countries combined).

For real news, a significant effect of condition on the pre-post difference in perceived manipulativeness is found (F(2,1774) = 42.73, p < 0.001,  $\eta^2 = 0.046$ , d = 0.44). A Tukey post-hoc comparison shows that the perceived manipulativeness of real news is significantly higher in the *Go Viral!* condition compared to the control condition (M = 0.51 vs M = 0.15,  $M_{diff} = 0.36$ , 95% CI [0.25, 0.46],  $p_{tukey} < 0.001$ , d = 0.45). However, the Infographics condition does not differ significantly from the control condition (M = 0.14 vs M = 0.15,  $M_{diff} = 0.01$ , 95% CI [-0.01, 0.11],  $p_{tukey} = 0.950$ , d = 0.02). These results are similar across countries; see on OSF, Supplementary Table S6. Thus, partial support is found for hypothesis H<sub>3</sub>: playing the *Go Viral!* game initially increases the perceived manipulativeness of fake news about COVID, but also of real news (although the effect size is descriptively smaller than for fake news). The UNESCO infographics, on the other hand, only show this effect for fake news (albeit to a lesser degree than *Go Viral!*), but not for real news.

To test hypothesis  $H_{7a}$  and  $H_{7b}$  for the manipulativeness measure, a repeated measures oneway ANOVA was conducted with condition as the between-subjects factor and time (pre – post – follow-up) as the within-subjects factor, for the UK sample (which was invited to the one-week followup). Doing so shows a significant effect of time \* condition on the perceived manipulativeness of fake news (F(4,1206) = 5.85, p < 0.001,  $\eta^2 = 0.004$ , d = 0.13). Specifically, in the one-week follow-up, UK participants in the *Go Viral!* condition rated fake news as significantly more manipulative than the control group (M = 5.96 vs M = 5.64,  $M_{diff} = 0.32$ , 95% CI [0.06, 0.59],  $p_{tukey} = 0.011$ , d = 0.30). The Infographics condition, however, did not differ significantly from the control group in the follow-up (M = 5.73 vs M = 5.64,  $M_{diff} = 0.09$ , 95% CI [-0.15, 0.32],  $p_{tukey} = 0.64$ , d = 0.08). As a robustness check, a between-subjects ANCOVA was conducted with pre-test manipulativeness as the covariate and manipulativeness in the follow-up as the dependent variable. Doing so shows a significant effect of condition on perceived manipulativeness (F(2,602) = 5.54, p = 0.004,  $\eta^2 = 0.013$ , d = 0.23), in that when controlling for the pre-test, participants in the *Go Viral!* condition find fake news significantly more manipulative in the follow-up than the control group ( $M_{diff} = 0.29$ , 95% CI [0.07, 0.51],  $p_{tukey} = 0.005$ , d = 0.27) and the Infographics condition ( $M_{diff} = 0.26$ , 95% CI [0.04, 0.49],  $p_{tukey} = 0.015$ , d = 0.25).

For real news, while a repeated measures between-subjects ANOVA shows a significant effect of time \* condition on the perceived manipulativeness of real news (F(4,1206) = 4.62, p = 0.001,  $\eta^2 = 0.003$ , d = 0.11), there was no significant difference between conditions for real news manipulativeness in the one-week follow-up (F(2,603) = 2.04, p = 0.131,  $\eta^2 = 0.007$ , d = 0.17), indicating that participants across conditions rated real news as equally manipulative in the follow-up study. In addition, a between-subjects ANCOVA with pre-test real news manipulativeness as the covariate and real news manipulativeness in the follow-up as the dependent variable gives no significant difference between conditions (F(2,602) = 2.35, p = 0.10,  $\eta^2 = 0.005$ , d = 0.14). Thus, the results provide support for hypothesis  $\mathbf{H}_{7a}$  and  $\mathbf{H}_{7b}$ : one week after the initial UK intervention, *Go Viral!* players continued to rate fake news about COVID-19 (i.e., fake news that they had not seen before) as significantly more manipulative than the control group and people who read the UNESCO infographics, whereas the initial scepticism of real news that was observed among *Go Viral!* players was no longer detectable one week after the intervention. Figure 22 shows these results in a bar graph.



Figure 22: Bar graphs of perceived manipulativeness of fake news and real news (UK only), by condition, for the pre-test (T1), post-test (T2) and 1-week follow-up (T3).

# Confidence

For the confidence measure, a between-subjects ANOVA on the pre-post difference in confidence scores for fake news is significant (F(2,1774) = 29.47, p < 0.001,  $\eta^2 = 0.032$ , d = 0.36), in that participants in the *Go Viral!* condition are significantly more confident after the intervention in their assessment of fake news than the control group (M = 0.34 vs M = 0.05,  $M_{diff} = 0.29$ , 95% CI [0.20, 0.38],  $p_{tukey} < 0.001$ , d = 0.44) and the Infographics condition (M = 0.34 vs M = 0.14,  $M_{diff} = 0.20$ , 95% CI [0.11, 0.29],  $p_{tukey} < 0.001$ , d = 0.29)<sup>27</sup>. In addition, participants in the Infographics condition are also significantly more confident in their assessment of fake news manipulativeness than the control group (M = 0.14 vs M = 0.05,  $M_{diff} = 0.09$ , 95% CI [0.002, 0.18],  $p_{tukey} = 0.043$ , d = 0.15). These results are similar (and significant) in all 3 countries, see on OSF Supplementary page, Table S8<sup>28</sup>.

For real news, a between-subjects ANOVA shows no significant difference between conditions for the pre-post difference in confidence scores (F(2,1774) = 1.58, p = 0.206,  $\eta^2 = 0.002$ , d = 0.09). This finding is similar in all 3 countries; see Supplementary Table S8. We thus find support for hypothesis **H**<sub>4</sub>: participants in both the *Go Viral*! and Infographics conditions become significantly more confident in their ability to assess the manipulativeness of fake news about COVID-19 but show no change for real news. These results are again similar in all 3 countries (see OSF page, supplementary analyses).

<sup>&</sup>lt;sup>27</sup> A reliability analysis shows good internal consistency for the confidence measure for the fake news items (M = 5.34, SD = 1.01, Cronbach's  $\alpha = 0.86$ ) and the real items (M = 4.89, SD = 1.15, Cronbach's  $\alpha = 0.90$ ), so we report the results for the average of both fake items and real items. See also Supplementary Table S9.

<sup>&</sup>lt;sup>28</sup> This study examined if participants become more confident in their assessment of fake news about COVID-19 for the right reason, i.e., they only indicate greater confidence if they also perceive fake news to be more manipulative. To do so, an ANOVA with the pre-post difference in fake news confidence as the dependent variable, and condition (*Go Viral!*, Infographics, control) and "manipulativeness-update" (a binary variable that is positive if the pre-post manipulativeness score for fake news is positive and negative if this difference is negative) as fixed factors was conducted. Doing so shows a significant effect of condition \* manipulativeness-update on fake news confidence, in that participants across conditions who see fake news as more manipulative after the intervention compared to before also show significantly greater confidence than participants who see fake news as *less* manipulative post-intervention compared to before. This effect is largest (descriptively) for *Go Viral!* participants. Thus, the findings suggest that that participants indeed update their confidence in the right direction; see Supplementary Tables S29-S30.

To test hypothesis  $H_{7a}$  and  $H_{7b}$  for the confidence measure, a repeated measures one-way ANOVA was conducted with condition as the between-subjects factor and time (pre - post - followup) as the within-subjects factor, which shows a significant effect of time \* condition on confidence to assess the manipulativeness of fake news (F(4, 1206) = 1.19, p = 0.009,  $n^2 = 0.003$ , d = 0.11). While a Tukey HSD post-hoc comparison gives no significant difference between conditions (all ps > 0.12), a between-subjects ANCOVA with pre-test confidence as the covariate and follow-up confidence as the dependent variable is significant (F(2,602) = 4.44, p = 0.012,  $\eta^2 = 0.010$ , d = 0.20), so that participants in the Go Viral! condition are significantly more confident than the control condition in their assessment of the manipulativeness of fake news one week after the intervention when controlling for the pre-test ( $M_{diff} = 0.25\ 95\%$  CI [0.05, 0.45],  $p_{tukey} = 0.011$ , d = 0.23). The results for the Infographics condition compared to the control group ( $p_{tukey} = 0.15$ ) and the Go Viral! condition compared to the Infographics condition ( $p_{tukey} = 0.45$ ) are not significant. For real news, a betweensubjects ANCOVA with pre-intervention confidence as the covariate and follow-up confidence as the dependent variable is significant (F2,602) = 3.49, p = 0.031,  $\eta^2 = 0.008$ , d = 0.18) so that Infographics participants are significantly more confident in their assessment of the manipulativeness of real news than Go Viral! participants ( $M_{diff} = 0.24, 95\%$  CI [0.004, 0.47],  $p_{tukey} = 0.045, d = 0.21$ ), whereas we find no difference between the Go Viral! and control condition ( $p_{tukey} = 0.85$ ) or the Infographics and the control condition ( $p_{tukey} = 0.09$ ). Thus, these results provide support for hypothesis  $H_{7a}$ : one week after the intervention, people who played Go Viral! remained significantly more confident in their ability to assess the manipulativeness of fake news, but not real news. In partial support of  $H_{7b}$ , participants who read the UNESCO infographics remained more confident in their ability to assess the manipulativeness of real news, but not fake news.

#### Sharing misinformation

For the sharing measure, a between-subjects ANOVA on the pre-post difference in willingness to share fake news is significant (F(2,1774) = 4.00, p = 0.019,  $\eta^2 = 0.004$ , d = 0.13), in that participants in the *Go Viral!* condition are significantly less likely to share fake news after the intervention than the control group (M = -0.18 vs M = -0.07,  $M_{diff} = 0.11$ , 95% CI [0.014, 0.19],  $p_{tukey} = 0.019$ , d = 0.15)<sup>29</sup>.

<sup>&</sup>lt;sup>29</sup> A reliability analysis shows good internal consistency for the sharing measure for both the fake news items (M = 2.34, SD = 1.31, Cronbach's  $\alpha = 0.91$ ) and the real news items (M = 3.17, SD = 1.25, Cronbach's  $\alpha = 0.89$ ), so we report the results <sup>for</sup> the average of both real and fake news items here. See also Supplementary Table S11.

However, we find no significant difference between the Infographics condition and the control group  $(p_{tukey} = 0.12)$  and the *Go Viral!* condition and the Infographics condition  $(p_{tukey} = 0.71)$ . These effects are directionally similar but not significant in each individual country (see Supplementary Table S10).

For real news, no significant pre-post difference for the sharing measure between conditions is found (F(2,1774) = 0.28, p = 0.75,  $\eta^2 = 0.0003$ , d = 0.03), with similar results across countries (see Supplementary Table S10). Thus, these results provide partial support for hypothesis **H**<sub>5</sub>: in the pooled sample, participants who played *Go Viral!* were significantly less willing than a control group to share fake news about COVID-19 with people in their network. However, no such effects for infographics participants are evident, and the effect is only visible in the higher-powered pooled sample (but not in country-level analyses). This may be the result of a floor effect given that average sharing intentions are generally low.

To test hypothesis  $\mathbf{H}_{7a}$  and  $\mathbf{H}_{7b}$  for the sharing measure, a repeated measures one-way ANOVA was conducted with condition as the between-subjects factor and time (pre – post – followup) as the within-subjects factor, which shows no significant effect of time \* condition on the selfreported willingness to share fake news (F(4,1206) = 0.94, p = 0.44,  $\eta^2 = 0.001$ , d = 0.06). In addition, a between-subjects ANCOVA with pre-test willingness to share fake news as the covariate and willingness to share fake news in the follow-up as the dependent variable gives no significant difference between conditions (F(2,602) = 1.50, p = 0.22,  $\eta^2 = 0.003$ , d = 0.11). Similarly, for real news, a repeated measures one-way ANOVA shows no significant effect of time \* condition on the self-reported willingness to share real news (F4,1206) = 0.99, p = 0.412,  $\eta^2 = 0.001$ , d = 0.06). Thus, the results provide no support for hypothesis  $\mathbf{H}_{7a}$  and support for  $\mathbf{H}_{7b}$  for the sharing measure; one week after the intervention, there is no longer a difference between conditions in terms of the selfreported willingness to share either real news or fake news about COVID-19.

#### Sharing the intervention

To test hypothesis  $H_6$ , an independent samples *t*-test with condition (*Go Viral!* or Infographics) as the independent variable and willingness to share the game or infographics on their social media accounts as the dependent variable was conducted. Indeed, the results suggest that *Go Viral!* participants are significantly more willing than Infographics participants to share the intervention with people in their network ( $M_{goviral} = 3.95$  vs  $M_{infographics} = 3.58$ ,  $M_{diff} = 0.37$ , 95% CI [0.15, 0.61], p = 0.001, d = 0.19, 95% CI [0.08, 0.31]), in support of hypothesis H<sub>6</sub>.

## Motivational threat

An ANOVA with traditional threat (e.g., fear, anxious, dangerous) as the dependent variable and experimental condition (*Go Viral!*, Infographics, control) and threat/post-test order as the independent variables revealed a non-significant effect (F(2,1774) = 2.46, p = 0.085). In contrast, an

ANOVA with motivational threat (e.g., "*I want to defend my current attitudes from misinformation about COVID-19*") as the dependent variable and experimental conditions and threat/post-test order as the independent variables, shows that the overall model is marginally significant (F(5,1771) = 2.19, p = 0.05). This is primarily driven by the condition the participants were in (F(2, 1771) = 4.06,  $p = 0.01, \eta^2 = 0.005$ ). Specifically, a Tukey HSD post-hoc comparison shows that, compared to the control condition, averaged motivational threat was significantly higher in the *Go Viral!* condition ( $M_{goviral} = 5.59 \text{ vs } M_{control} = 5.44, M_{diff} = 0.15, p_{tukey} = 0.05, 95\%$  CI [0.002, 0.29], d = 0.14). Similarly, motivational threat in the *Go Viral!* condition was significantly higher compared to the Infographics condition ( $M_{goviral} = 5.59 \text{ vs } M_{infographics} = 5.43, M_{diff} = 0.158, p_{tukey} = 0.03, 95\%$  CI [0.01, 0.30], d = 0.15. There was no significant difference between the control and Infographics condition ( $p_{tukey} = 0.97$ ). These results support the exploratory hypothesis **H**<sub>8</sub>.

# Exploratory analyses

In order to examine whether findings in Chapter 2 and Chapter 3 are consistent with the large and cross-cultural data-set in the present data, additional mediation analyses were run to examine the role of attitude certainty in the inoculation process against online misinformation.

# Role of confidence in sharing intentions

To investigate the role of attitude certainty on sharing intentions, a simple mediation analysis was performed. First, the independent variable consisted of the experimental condition (control, inoculation), the outcome variable was the sharing intentions of news items (fake/real), and the mediator variable for the analysis was confidence (pre-post).

The results revealed that the total effect of the treatment conditions on perceived manipulativeness of fake news items was significant,  $\beta = 0.05 \ t = 2.75$ , p < .006, 95% CI(0.02, 0.09). With the inclusion of the mediating variable (confidence in the assessment of fake items), the impact of condition on the perceived manipulativeness of fake news items was still found significant,  $\beta = 0.43$ ,  $t=2.26 \ p=.02$ , CI(0.005,0.08). The indirect effect of the treatment conditions on sharing of fake items through confidence was also significant,  $\beta = 0.008$ , t= 2.5, p = .01, CI(0.002, 0.02). This shows that the relationship between condition and sharing of misinformation items is partially mediated (17.0%) by participants' confidence in their reliability assessment. Thus, these findings provide support for attitude certainty playing a mediating role in sharing intentions of fake news items (Figure 23). Conversely, results suggest no total, direct, or indirect effect for real news.  $\beta = 0.11$ , t= 0.5, p < .72, 95% CI (-0.03, 0.05).



Figure 23: Path plot of mediation analysis on the relationship between condition and sharing intentions as mediated by confidence. \*p <.05.

# Rollout campaign of (psychological) vaccine

Lastly, insights gained through Google Analytics facilitate the tracing of the inoculation intervention since its launch. Below, the distribution of *Go Viral!* is depicted by the top five countries and at the global level. Interestingly, the data also suggests that, since its launch, the game was accessed more than 50,000 times through a "challenge your friend" referral and that, when combined with access via social media, it accounts for 22.1% of all "vaccine uptake". These results further emphasise the potential for psychological herd immunity against misinformation by having inoculated individuals pass on the treatment.



Figure 24: Distribution of Go Viral! since launch in October 2020.

## Robustness checks

A series of linear regressions were conducted with average fake news and real news manipulativeness, confidence, and willingness to share as dependent variables, and condition along with several covariates (gender, age, political ideology, whether participants had COVID-19, their level of worry

about COVID-19, their COVID-19 vaccine intentions, and whether the threat measure or item-rating task was presented first in the post-test) as independent variables. Similar estimates were returned when controlling for various covariates at the pooled and individual country level (e.g., gender, age, political identity; Supplementary Tables S16-S23). The study finds that all effects reported above are robust and that there is no consistent interaction with any of the dependent variables (manipulativeness, confidence, and sharing of real and fake news) for any of the covariates mentioned above, indicating that the effects conferred by both interventions are similar across different demographics and attitudes about COVID-19.

Finally, the study checked whether participants' self-reported level of vigilance with respect to information content and news sources (Newman *et al.*, 2020) was related to fake news manipulativeness assessments. A linear regression shows that while higher vigilance is significantly associated with higher perceived manipulativeness of fake news in the pre-test (i.e., before the intervention;  $\beta = 0.15$ , 95% CI [0.10, 0.19], p < 0.001), this effect is no longer significant across the 3 conditions for the pre-post difference in fake news manipulativeness ( $\beta = 0.04$ , 95% CI [-0.0039, 0.087], p = 0.07). This indicates that higher concern about the accuracy and source reliability of one's news consumption is associated with lower susceptibility to COVID-19 misinformation, but not with the learning effect conferred by the inoculation treatment or the infographics.

# 4.9 Discussion – Study 2

Study 2 demonstrates that both prebunking interventions, the *Go Viral*! game and the UNESCO infographics, significantly increase the perceived manipulativeness of fake news about COVID-19, compared to a control group. This result is in line with Study 1 and remained valid in a randomised controlled setting and across 3 different languages. *GoViral*! participants were also significantly more confident in their judgments and experienced more motivational threat to defend their attitudes. With regards to real news, unlike Study 1, we find ambiguous results in Study 2. We find that the *Go Viral*! game increases people's scepticism of real news immediately after the intervention (similar to findings by Guess *et al.*, 2020), whereas this effect is not observed for the UNESCO infographics. However, scepticism of real news among *Go Viral*! participants dissipate entirely after one week, with real news

being rated as equally manipulative across conditions in the follow-up<sup>30</sup>, while participants who played the game continued to rate fake news as more manipulative in our longitudinal study. We found no differences between conditions for real news for the confidence and sharing measures. Consistent with the Chapters 2 and 3, this study replicated the finding that attitude certainty mediates sharing intentions of misinformation. Having designed and tested three separate games (*Bad News, Join this Group, Go Viral*) which each concerned themselves with different topics and forms of misinformation, the accumulated findings further underline the efficacy of active inoculation efforts spanning across topics, platforms, and potentially individuals.

# 4.10 General Discussion

Across two large-sample studies using different research designs, this chapter finds strong support that both active and passive prebunking interventions increase people's ability to spot manipulative social media content related to COVID-19. In addition, and in line with previous studies (Basol, Roozenbeek and van der Linden, 2020; Roozenbeek and van der Linden, 2020), the results show that prebunking increases people's confidence in their ability to spot manipulative content. Crucially, this certainty update occurs in the right direction, in the sense that people only become more confident of the degree to which they believe a certain social media post is manipulative when they also assess the manipulativeness of this post correctly after the intervention. Furthermore, the results are consistent with previous findings in that they further emphasise the mediating role of attitude certainty in the process of building, strengthening, and spreading resistance. For Go Viral! players, these two effects remain significant for at least one week after gameplay, even when presented with fake news about COVID-19 that they have never seen before, indicating robust support for a high degree of retention of the inoculation effect (Maertens et al., 2020). These results also speak to the relative benefit of active versus passive inoculation, especially in terms of preventing decay of the effects over time. Furthermore, it is noteworthy that, at least descriptively, the active intervention yielded larger effect sizes for manipulativeness and confidence assessments than the passive intervention.

 $<sup>^{30}</sup>$  We note that the real news items at follow-up were different items (to avoid item-memorisation effects).

# Boosting attitude certainty through inoculation

Within the context of attitude certainty, these findings contribute to the current limited understanding of the role of certainty in the inoculation process. Considering research on persuasion which suggests that the conviction with which an attitude is held predicts behavioural intentions, strengthens resistance to persuasion, persists over time, and leads individuals to talk to others about their opinion more (Tormala & Petty, 2004). This chapter provides additional support for the role of attitude certainty in generating and strengthening inoculation-based resistance to misinformation but also spreading resistance toward a blanket of protection against differing misinformation techniques. Next, research on generalised and prophylactic inoculation treatments should examine how the confidence boost can be leveraged to spread resistance from the inoculated individual to their social network.

# Stopping the spread of misinformation

With respect to people's willingness to share social media content about COVID-19, the findings show that playing the Go Viral! game significantly reduces willingness to share misinformation about the virus (in line with Roozenbeek and van der Linden, 2020). However, this finding is not significant at the country level (although directionally similar). Furthermore, this effect was no longer significant after one week, and no difference in sharing willingness for the UNESCO infographics was found. As such, the #ThinkBeforeSharing infographics finding is inconsistent with recent research showing that getting people to pause and think can help reduce sharing of false news online (Fazio, 2020). It is possible that flooring effects are at play in our study (i.e., participants had relatively low willingness to share both fake and real news even in the pre-test). Another possibility is that the sample sizes for the individual countries were not large enough to detect a significant effect; for example, a post-hoc power analysis for the sharing measure with d = 0.15,  $\alpha = 0.05$  and n = 710(which is what we obtained for the UK sample), returns an achieved power of 0.41. It is thus possible that larger samples are needed to find consistent effects of misinformation interventions on sharing intentions. For example, other recent interventions aimed at reducing the spread of COVID-19 misinformation, such as accuracy nudges, have also shown relatively small effects on sharing intentions (Pennycook et al., 2020). Regardless, both Chapters 3 and 4 provide support for active inoculation interventions affecting sharing intentions of misinformation. Next, research should explore ways to explore whether inoculated individuals do actually spread less misinformation after game-play.

# The state of Truth

This study also adds to the ongoing debate about the extent to which anti-misinformation interventions influence people's assessment of real news (Guess *et al.*, 2020; Maertens *et al.*, 2020; Pennycook *et al.*, 2020). The findings are somewhat ambiguous: while in Study 1 playing the *Go Viral!* game does

not meaningfully affect people's assessment of real news, the findings in Study 2 suggest that Go Viral! players find real news about COVID-19 significantly more manipulative immediately after gameplay. Curiously, this effect is observed in Study 2 even for the items that were used in both studies. At the same time, confidence assessments and sharing intentions of real news are not affected by prebunking interventions (unlike for fake news), and any heightened scepticism of real news dissipates entirely one week after playing (whereas Go Viral! players continue to rate fake news as significantly more manipulative than the control and Infographics groups). These results may be put into perspective with the decline of trust in news in recent years (Newman et al., 2019, 2020). For instance, a recent study found that across four countries (Germany, Spain, the United Kingdom, and the United States), internet users' navigation on social media was based on a "generalised scepticism" (Fletcher and Nielsen, 2018). Overall, our findings suggest that while prebunking interventions may (sometimes) influence people's assessment of real news (see also Guess et al., 2020), the presence and size of this effect varies substantially across studies and designs, and in the absence of established psychometric scales could be due to item-effects rather than genuine underlying scepticism (Roozenbeek, Maertens, et al., 2020). Thus, further exploration and theorising on the implications of heightened scepticism of real versus fake news for overall improvements in veracity discernment is needed. For example, simple fact-checking is not without risk either (in some cases it can backfire, Peter and Koch, 2015; Ecker, Lewandowsky and Chadwick, 2020), so the question for any intervention is whether the benefits outweigh any potential side effects.

### Motivational threat as a key component of inoculation

The notion of threat being integral in conferring attitudinal resistance has long played a central part in inoculation theory, with scholars arguing that "inoculation would be impossible without threat" (Compton and Pfau, 2005: 100–101). This observation is interesting because McGuire never explicitly measured threat himself, and a meta-analysis showed no significant relationship between threat and resistance, urging inoculation researchers to take a closer look at the role of threat (Banas and Rains, 2010). More recent scholarship suggests that "motivational threat"—or threat in the form of motivation defend oneself against persuasive attacks—is conceptually more consistent with inoculation-generated resistance (Banas and Richards, 2017; Richards and Banas, 2018). Our findings add to this debate by demonstrating a clear main effect of motivational threat (but not apprehensive threat) of *GoViral!* on the perceived manipulativeness of fake news. Additionally, the results also suggest that only motivational threat (not apprehensive threat) is a significant predictor of perceived manipulativeness, attitude certainty, and sharing of real news. At the theoretical level, two main conclusions can be drawn from this.

First, these results are consistent with Banas and Richards (2017), emphasising that the traditional operationalisation of apprehensive threat is not adequately capturing the threat elicited through an inoculation treatment and that, instead, approaching threat as a motivation offers a better representation of the psychological process underpinning resistance through inoculation. Indeed, the

crucial difference between these two measures of threat is the element of fear and the present findings support the notion that reliance on fear-based items has contributed to threat's lack of predictive validity of resistance (Banas & Rains, 2010). Secondly, the results also shed initial light on the potential interaction between attitude certainty and motivational threat. While more research is needed to further dissect and understand the specific dynamics within the inoculation process, these results revealed that motivational threat is a superior mediator and functions as a mechanism by which inoculation confers resistance.

Consistent with O'Keefe (2003), these results suggest that psychological states during the inoculation process should be treated as "potential mediators of persuasive effects" (p.269). In fact, more than two decades ago, Pfau (1997) advised inoculation scholars that "efforts to enrich inoculation theory require going back to the construct's core assumptions, refining and extending them, and then testing the reformulated logic in laboratory studies" (pp.151-152). The present chapter aimed to do that by revisiting threat, one of the theoretical pillars, and assessing the renewed conceptualisation put forward by Banas and Richards (2017) in a large-scale, randomised, longitudinal, and cross-cultural setting. Consequently, at the practical level, this suggests that *Go Viral!* is a promising step toward interventions that, even when applied to highly unprecedented and uncertain topics (e.g., pandemic), *motivate* people to engage in attitude-bolstering resistance processes without inadvertently heightening anxiety around the imminent attack or vulnerability of one's attitudes.

# Rolling out the psychological vaccination

Lastly, consistent with the role that attitude certainty appears to play in the strengthening and spreading of attitudinal resistance across differing, previously unseen misinformation items and techniques, research should also begin exploring whether inoculation can be spread at an interindividual level. On the one hand, insights gained through Google Analytics demonstrate a wide distribution of Go Viral! across all continents (many accessed through a "challenge your friend" link), emphasising the scalability of active, therapeutic, and generalised inoculation treatments against misinformation. Indeed, Lewandowsky and van der Linden (2021) argued that the future of inoculation research will be best served if it focuses on how inoculation may spread from one person to another and starts offering realistic predictions about building psychological "herd immunity" against the increasing spread of online misinformation. Arguably, this constitutes a first pathway towards psychological herd immunity. Namely, by designing and testing a generalised intervention that is effective against a variety of (previously unseen) arguments, across different cultures, and with an incentive to outperform the score of the person who challenged you, Go Viral! is arguably offering promising directions towards a society that is collectively able to spot and resist misinformation. Next, and perhaps theoretically more interesting avenue concerns the question whether *vicariously* receiving a diluted form of a treatment from a previously inoculated individual can confer resistance? If so, for how long could that "ripple effect" of inoculation persist?

# Shortcomings and future directions

Like any research, this research has several limitations. First, we were only able to conduct a oneweek follow-up in the UK. And while research on the decay of the inoculation suggests that, with regular "boosters", the effect persists for at least three months (Maertens et al., 2020), further research is needed to understand the efficacy of Go Viral! simultaneous to the societal and medical developments concerning the pandemic. In other words, just as booster shots are required to circumvent the continuous emergence of variants and mutations of the virus, it remains to be tested how generalisable and effective Go Viral! is against COVID-19 misinformation that perpetually changes in content and modality. Second, it should be noted that only configural or weak invariance across the three language versions of this survey was established for the manipulativeness, confidence, and sharing measures in Study 2 (both real and fake news items; see Supplementary Tables S24-S28 for invariance testing). However, the alpha levels for each construct were acceptable at the pooled and individual country level (0.72 - 0.92); see Supplementary Table S5). It has been suggested that a failure to reach a certain level of invariance should not prevent analyses so long as researchers note this limitation, which I do here (Putnick and Bornstein, 2016). Furthermore, as mentioned above, the study may have not achieved enough power to detect differences between conditions for the sharing measure, and although judgments were made within a simulated social media setting (enhancing ecological validity), we note the self-reported nature of our data. Third, in selecting the treatment comparison (UNESCO's infographics), real-world generalisability was prioritised. Considering that a large portion of efforts regarding scientific communication and debunking myths about the virus during the pandemic heavily relied on infographics, texts, and elaborate factual content, the present study takes a critical look at predominant efforts against health misinformation. Indeed, this is a call for governments and relevant institutions to anchor pre-emptive debunking methods and other psychological insights in the centre of their designs when aiming to fight harmful misinformation. Regardless, selecting a real-world intervention limited the internal validity for this study so future research may want to adopt a passive control that is more balanced and identical to the active condition across key parameters (e.g., other gamified interventions).

## 4.11 Conclusion

Across two large-sample studies, we provide strong cross-cultural evidence for the effectiveness of two short and scalable prebunking interventions as a tool to reduce susceptibility to misinformation about COVID-19. One of these interventions, a 5-minute free-to-play browser game, has a demonstrable positive impact on people's ability to identify fake news about the virus for at least one week after playing. Thus, we suggest that prebunking constitutes a crucial step toward combating harmful misinformation about the pandemic. Lastly, as the success of COVID-19 vaccination programs worldwide depend in part on minimising the amount of unreliable information that surrounds them, our findings add to the emerging insight that behavioural science is a crucial tool to

help mitigate the spread of misinformation online. To do so, inoculation research must begin to offer realistic predictions for conferring psychological herd immunity against misinformation. This chapter presents one potential pathway by highlighting the efficacy of *Go Viral!*, an inoculation treatment I helped develop, test, and launch cross-culturally. By collaborating with the UK Cabinet Office and implementing *Go Viral!* in WHO's campaign against COVID-19 misinformation, this chapter further represents a decisive commitment to leverage inoculation theory in the multi-disciplinary fight against misinformation. Next, reviewing and challenging the boundary conditions of inoculation theory will allow this doctoral thesis to establish whether and how inoculation-based resistance may be spread between individuals.

5 HARNESSING POST-INOCULATION TALK TO CONFER INTRA- AND INTERINDIVIDUAL RESISTANCE TO PERSUASION

# 5.1 Abstract

For decades, inoculation theory has regarded counterarguing, one of its core theoretical concepts, as an intrapersonal and subvocal process to attitudinal resistance. More recently, the focus has shifted to vocalised and interpersonal counterarguing taking form through post-inoculation talk (PIT). Across three phases, two preregistered (n = 400, and n = 600), this chapter set out to address three theoretical gaps in the inoculation literature, (1) whether post-inoculation talk occurs organically, (2) whether engaging in it has any effects on the spreader, and (3) whether receiving PIT can vicariously confer resistance to its recipients. To do so, participants first received a text-based inoculation treatment against a fictitious chemical and required individuals to leave an *intended* pass-along message. A week later, the follow-up study investigated whether actual inoculation talk had occurred, what it was about, and whether engaging in PIT, in turn, affected the spreader's resistance process. Lastly, the actual PIT messages composed during the follow-up study were used in place of traditional inoculation treatment messages to inoculate a new sample against misinformation about the fictitious chemical. Overall, the findings suggest that inoculated individuals do voluntarily engage in post-inoculation talk without any prior instructions, and that PIT plays a role in predicting perceived resistance, strengthening attitudes, and bolstering attitude certainty. Lastly, the findings suggest that PIT messages can vicariously inoculate the recipients of talk. In establishing these effects, the current chapter advocates for an updated conceptualisation and application of PIT. More specifically, by treating it first as a form of vocalised counterarguing, then a process of resistance through inoculation, and lastly, as a product of inoculation - this chapter points towards a path on which leveraging PIT could lead toward psychological herd immunity against online misinformation. Limitations and future directions are discussed.
# 5.2 General introduction

While the previous chapter critically revisited one theoretical key component (threat) and identified one potential pathway towards psychological herd immunity against misinformation via interpersonal sharing of the inoculation intervention on social media, this chapter aims to review the theoretical counterpart (counterarguing) and assess whether post-inoculation talk can offer a second path. After all, although Go Viral! has experienced an impressive uptake (over one million 'shots'), if inoculation is to keep up with, if not outpace misinformation, its efforts cannot rely on having every single person play a game (and regularly "get boosted"). Instead, this chapter aims to examine whether inoculation could possibly be passed on from one person to another by talking about it. The previous chapters have established that inoculated individuals become more confident and more willing to share the treatment with others (and crucially, less misinformation). In light of previous research highlighting that confidently inoculated individuals tend to speak out about the issue topic more (Lin & Pfau, 2007; Tormala & Petty, 2004), the next step then concerns itself with getting individuals to talk more about the inoculation treatment itself. Put simply, now that I have identified both the confidence-boosting effect as well as the motivating impact of threat on the inoculation process, can these accumulated insights be applied in a way that leads inoculation to be passed on from one person to another? To answer this, the theoretical foundations of counterarguing must be revisited first.

# Revisiting counterarguing

Inoculation theory builds from work on message-sidedness (Lumsdaine & Janis, 1953) and proposes a method of attitudinal resistance analogous to the immunisation process (McGuire, 1964). The notion is that raising and refuting attitude-challenging arguments confers resistance to future stronger challenges by (1) explicitly forewarning (threat) the individual and thereby, revealing a vulnerability of having their attitudes attacked, which, in turn, (2) motivates the generation of *additional* counterarguments to those provided in the pre-treatment). (Compton, 2013) has referred to these two key components as a catalyst (threat) and an activity (counterarguing) within the process of inoculation-conferred resistance. Counterarguing describes the generation of counterarguments and refutations in the interim between exposure to the inoculation treatment and the subsequent attack message (Compton, 2013). While a pre-treatment message does usually present a counterargument and their respective refutations in a two-sided approach, *the process* of counterarguing is argued to be a dynamic activity where inoculated individuals begin to raise and refute arguments about the issue topic on their own (Compton & Pfau, 2005). This process is closely tied to the medical analogy of inoculation theory, suggesting that refutations act much like antibodies, attacking and weakening the antigens (McGuire, 1964; Compton, 2013).

Though a large body of inoculation research confirmed that information-induced resistance to persuasion is effective, a lack of empirical confirmation as to *how* inoculation conferred resistance was evident in the early years of the theory (Ivanov et al., 2018). Indeed, both threat and counterarguing were mostly assumed (Compton & Pfau, 2005). Since then, inoculation scholarship has experienced a resurgence, demonstrating its efficacy across a variety of contexts (for a series of meta-analyses and reviews on the efficacy of inoculation, see Banas & Rains, 2010; van der Linden et al., 2021; Compton et al., 2020). Although most of the inoculation research has focused on understanding the processes of cognitive resistance, perhaps one of the more intriguing recent avenues concerns itself with the potential to spread and diffuse resistance over populations not initially exposed to the inoculation messages (Basol et al., 2021; Compton et al., 2021). There is reason to believe that such spreading of inter-individual resistance through inoculation is possible. To begin with, studies have demonstrated that inoculation treatments increase issue- involvement in the target topic (Compton & Pfau, 2009) as well as the willingness to talk about societally contested issues (Lin & Pfau, 2007). This suggests that the process of counterarguing might play a larger role than so far captured by inoculation research.

# Ripple effect of counterarguing: post-inoculation talk

Inoculation theory posits counterarguing to be a crucial pillar to conferring attitudinal resistance. Interestingly, the process of counterarguing has traditionally been exclusively conceptualised as a subvocal process, disregarding the impact of fully vocalised counterarguing, through interpersonal talk (Compton & Pfau, 2009). In other words, rather than considering the role of counterarguing through actual talk, counterarguing has predominantly been operationalised as the inoculated individuals' internal dialogue with the anticipated attack message. Consequently, most inoculation studies have limited their assessment of subvocal counterarguing (i.e., internal dialogue) outputs to thought listing processes (Brock, 1967) and recognition check-off tasks (Pfau et al., 2004; Ivanov & Pfau, 2012). Hence, counterarguing was typically measured by participants either actively listing counterarguments against the attack message themselves or identifying and rating a list of arguments alongside their own. More recently, Ivanov and colleagues (2012) conducted a three-phased study to directly assess counterarguing processes in inoculation, whether subvocal, vocal or both. Their main finding suggested that inoculation generates more talk and that, in turn, more talk contributes to greater resistance. Indeed, Ivanov and colleagues (2012, p.704) have asserted that "counterarguing may be both intrapersonal and interpersonal, both subvocal and vocal", suggesting a more interconnected dynamic between different forms of counterarguing.

Subsequently, this raises critical questions regarding the theoretical conceptualisation of inoculation theory. That is if post-inoculation talk (PIT) plays a crucial role in the inoculation process, how does it fit into the analogical foundations of inoculation theory? Some scholars argued that PIT fits the analogy both pragmatically and conceptually, by comparing it to a metaphorical syringe, carrying and injecting the immunisation for producing resistance (Compton & Pfau, 2009; Parker et al., 2012). Like antibodies attacking a virus, some have regarded PIT as an antigen triggering the counterarguing response against an imminent threat (Ivanov et al., 2015). This pass-along effect can

play a significant role in sustaining the influence of inoculation (Ballew et al., 2019) if not even be as effective as the treatment itself (Southwell & Yzer, 2009). However, what the existing literature on PIT offers in breadth, it lacks in specificity. In other words, the current scientific understanding of PIT's criteria, benefits, and boundaries remains a mosaic, making accurate theoretical mapping difficult. When the logic of inoculation theory is challenged (Wood, 2007), most research continues to follow McGuire's (1964) use of the biological analogy as an explanatory force when formulating novel theoretical developments (Compton, 2013). Of course, as suggested by Compton, the analogical foundations of the theory should be regarded as more instructive than prescriptive, proposing that some phenomena in inoculation-induced resistance may be more abstract than so far suggested. In the case of accurately positioning post-inoculation talk within the inoculation process, therefore, inoculation scholarship risks reinforcing a simplified and false dichotomy by falsely separating internal and external communication (Anderson, 1967). Put plainly, reducing counterarguing to a single form of vocalisation fails to account for the multitudes in which counterarguing might occur (e.g., subvocal, vocal, imagined, online talk).

# What are the underlying mechanisms of PIT?

The limited research on PIT makes it difficult to disentangle the various factors involved in the (sub)vocal counterarguing process through inoculation. For instance, research can be split into two approaches to why PIT may occur in the first place. Compton and Pfau (2009) distinguish between reassurance talk and advocacy talk, suggesting that post-inoculation talk may be driven by different factors such as threat, confidence, affect, and attitude strength. More specifically, it is argued that threat and its resulting awareness of one's attitudinal vulnerability, though crucial to the inoculation process, are antithetical to confidence (e.g., Ivanov et al., 2012). The negative relationship between confidence and threat observed by Pfau and colleagues (2004) suggests that inoculated individuals are more likely to turn to others for reassurance after having had their attitudes shaken. Though this heightened vulnerability may be motivating reassurance-driven external talk, scholars have argued that it simultaneously, and not necessarily consciously, contributes to the spread of inoculationrelevant content across their social networks. Thus, while the Chapters 2 and 3 outlined the role of attitude certainty in resistance to persuasion, more research is needed to understand the role of confidence in the occurrence of post-inoculation talk. Additionally, Compton and Pfau (2009) suggest that other motivations may be at play on the other end of the confidence spectrum. That is, while threat reduces confidence, the refutational pre-emption component and generation of counterarguments likely enhance confidence. This suggests that there might be an interactive and somewhat reciprocal exchange where one's acute awareness of their vulnerability increases the motivation with which one equips themselves with counterarguments. This attitude-bolstering process increases attitude certainty, allowing inoculated participants to advocate for their beliefs, even if it is a minority opinion and contested issues (Lin & Pfau, 2007). This notion of inoculation acting to promote advocacy is further supported by the increase in issue involvement upon exposure to the treatment message (Pfau et al.,

2004). Considering that issue involvement seems to be positively correlated with Word-of-Mouth-Communications (WOMC) (Compton & Pfau, 2009; Giese et al., 2020), it can be hypothesised that the fluctuations of confidence in relation to attitude strength play an integral part in whether or not inoculated individuals decide to spread the treatment in their networks.

# Why does it matter?

Limited research suggests that word-of-mouth communication could be as effective as the direct receipt of a message and that it is likely to inspire more talk leading to relevant actions (Compton & Pfau, 2009). Thus, researchers have argued that post-inoculation talk (PIT) is likely to have similarly significant effects on the process and impact of inoculation. Moreover, findings show that PIT plays an active role in inoculation by increasing the likelihood of talk (Lin & Pfau, 2007), changing intentions about how they wish to talk about an issue (Compton & Pfau, 2004) and that increased talk about the treatment issue enhances resistance to persuasion (Ivanov et al., 2012). Lastly, Ivanov and colleagues (2015) find that inoculated individuals partake in increased advocacy attempts about the treatment issue, suggesting that PIT could serve beyond providing mere assurance and play a critical role in the spread of attitudinal resistance instead. Consequently, scholars assert that rather than viewing PIT as a mere product of inoculation, it is likely to play an integral part in the process itself. Lin & Pfau (2007) found that inoculated individuals were more likely to speak up about their attitudinal position – even in the face of anticipated disagreement. This proposes an important shift in the way inoculation theory is conceptualised. That is, by acknowledging and incorporating PIT as a key element of the inoculation process, its theorisation moves away from a solely intrapersonal process of resistance towards something more dialogic, more collective. This notion of the dynamic interplay between talk and resistance, where "inoculation generates more talk, and talk contributes to resistance" (Ivanov et al., 2012), is a scarcely examined avenue of how talk might play an integral part in strengthening and spreading attitudinal resistance within and between individuals.

# What questions remain unanswered?

Recently, scholars have called for inoculation research to provide actionable and realistic steps toward spreading attitudinal resistance from one person to another (Basol et al., 2021; Compton et al., 2021; Lewandowsky & van der Linden, 2021; van der Linden, 2022). As discussed in Chapter 3, research on the psychology of rumour suggests that rumours and gossiping are driven by a collective attempt at sense-making (Bordia & DiFonzo, 2017; DiFonzo & Bordia, 2011). Could such attempts be leveraged to spread prebunking strategies as well? Moreover, given that such processes are further reinforced and accelerated by social media, it is important to consider research on pass-along paradigms and communication chains (Giese et al., 2020) which demonstrate the increasing distortion of passed-on messages (i.e., "Telephone game" effect). How then could resistance against misinformation be passed on?

This doctoral thesis argued that counterarguing might offer a promising puzzle piece to the conundrum of spreading resistance against misinformation from one person to another. Thus far, the chapters have provided consistent support for the role of attitude certainty in building, strengthening, and spreading inoculation-based resistance. Additionally, revisiting threat, one of the two theoretical key components of inoculation theory, has rendered support for the need to rethink and measure threat as a motivation. Equally, by revisiting the other theoretical pillar, counterarguing, and examining the, until previously neglected, component of vocalised counterarguing (i.e., talk), this chapter aims to explore the possibility of spreading attitudinal resistance through post-inoculation talk. However, several questions remain unanswered that hinder the use of PIT to its full potential. For postinoculation talk to offer a viable solution toward spreading inoculation against misinformation, a series of assumptions must be tested and established first. To begin with, does post-inoculation talk occur organically? That is, contrasting previous research which provided specific directions to their participants, do inoculated individuals voluntarily talk to others about the inoculation treatment? partake in PIT? If so, how many conversations do they have? Does the frequency of conversations or conversation partners have an attitude-bolstering effect on the inoculated individual? In other words, does engaging in post-inoculation talk have any effects on the attitude (e.g., strength, confidence)? Of course, a more important question concerns the content of PIT. Especially when pitted against persuasive forms of misinformation, what do inoculated individuals talk about? Does the content of their talk differ from those not inoculated?

Once these pre-requisites and effects of PIT are established, inoculation research can then begin to raise the question of whether receiving a vicarious form of an inoculation treatment (i.e., receiving PIT from an inoculated individual) confers resistance against misinformation. Furthermore, research has yet to identify the prerequisites for verbally passed-on inoculation messages to be effective (e.g., whether PIT needs to mimic core components of traditional inoculation messages). Then, inoculation research can begin exploring how far "the ripples of resistance" extend (i.e., how many times PIT can be passed-on before it stops conferring resistance). In short, what role is PIT currently playing in the generation and strengthening of resistance that current inoculation research may not be aware of? And more importantly, what role can PIT potentially play in the passing on of attitudinal resistance from one person to another? Especially within the context of misinformation and the novel challenges posed by technological advances and its pervasive information infrastructures, exploring and optimising means to outpace the spread of misinformation is critical. This doctoral thesis proposes that inoculation scholarship can lead to two pathways towards psychological herd immunity. While Chapter 4 hinted at the first path of developing inoculation interventions to confer generalised resistance across cultures and designed specifically for its players to spread the "vaccine", this chapter aims to focus on the path of post-inoculation talk and its potential function as a catalyst resistance.

# Aims of this chapter

To summarise, this chapter aims to examine whether post-inoculation talk occurs organically, whether and how it impacts the individual spreading the inoculation via talk, and lastly, whether being 'vicariously inoculated', that is, receiving PIT, confers resistance to the recipient of PIT. To my knowledge, this is the first study extending the effects of PIT in a way that spans from an intraindividual focus to interindividual resistance and thereby situating it within the context of building psychological herd immunity. To do so, a three-phase inoculation experiment was conducted in which I aim to first establish whether participants do voluntarily engage in post-inoculation talk when no instructions are provided (Phase 1). Here, participants will also indicate their intended pass-along message to others. Building up on that, Phase 2 will present the follow-up study conducted a week after treatment exposure and will examine whether inoculated individuals did indeed voluntarily engage in talk in the interim and, more importantly what the recall of their actual conversations look like. Lastly, actual pass-along messages composed during Phase 2 will be used to develop secondorder inoculation treatments. That is, Phase 3 will aim to assess the efficacy of post-inoculation messages in vicariously inoculating the prospective recipients. Doing so will allow the present chapter to take a substantial step towards uncovering the potential of PIT in the efforts toward herd immunity against misinformation.

# PHASE 1

# 5.3 Establishing post-inoculation talk

In Phase 1, the main aim was to establish whether post-inoculation talk occurs organically after treatment exposure. Given that the limited research on PIT has mostly relied on instructing their participants whether or not to engage in PIT (Ivanov et al., 2012; 2016), examining whether, how much, and what individuals would voluntarily share with others post-treatment exposure constitutes a critical building block for understanding and harnessing PIT effectively. Thus, in Phase 1, I focus on *intended* pass-along messages first.

### Issue selection

Contrary to previous research on PIT (e.g., Ivanov et al., 2015; Pfau et al., 2005), the present study did not employ inoculation treatments on the traditional issues (i.e., legalisation of the sale and use of marijuana; government restriction of violent content on TV programs, banning the manufacture, sale, and possession of handguns, and legalisation of gambling). It has been argued that, for inoculation research to be conducted, two criteria need to be met: participants should have a firm attitude or position in place at the time of the treatment and issues should provide room for participants to hold their position either in support of or opposition to the issue topics (Pfau et al., 2009). Additionally,

although the existence of an attitude (not its valence nor extremity) is argued to allow "a threat of change" (Compton, 2019, p.333), recent research on differing pre-existing attitudes has questioned this assumption, demonstrating that inoculation treatments are not contextually bound (Wood, 2007; Compton et al., 2021; Basol et al., 2021).

Hence, this present study concerned itself with a previously unheard of, and in fact, an entirely fictitious chemical which was called tryptostine (a combination of different chemical names). The rationale behind this was two-fold – as the central aim of this study was to examine post-inoculation talk when pitted against misinformation, there were concerns about the consequences of having individuals share real misinformation about an existing chemical with others in their social network. Equally, to avoid any memory confounds, the attack message spreading misinformation was on a fictitious chemical which the sample could not have pre-existing attitudes towards. In doing so, this study aimed to take a closer look at the effects of inoculation treatments on attitude change and strength without issue involvement mediating it. However, to mimic the environment of the pandemic where health (mis)information is perceived through the lens of risk, focusing on the hypothetical dangers of the chemical aimed to induce a sense of urgency and danger without contributing to the spread of harmful misinformation.

### Methods

# Design

This pre-registered study implemented a between-subject design with the independent variable being treatment conditions (inoculation, control) and post-inoculation talk as the dependent variable. Additional covariates consist of pre-post test measures (attitude change, attitude certainty, attitude strength, motivational threat, apprehensive threat, perceived resistance, and counterarguing) through semantic differential items.

## Material

The survey was developed on Qualtrics and entailed time measurements (allowing for the review of the time spent on each question). All materials (including consent and debrief form) were displayed within that survey and accessed through the online crowd-sourcing platform *Prolific Academic*.

# Procedure

Individuals accessed the online survey on *Prolific Academic*, which has been argued to provide qualitatively enhanced data compared to the online research platform MTurk (Peer et al., 2017). Participants were required to read an information sheet and provide electronic consent before they could proceed with their participation. To minimize the risk of influential biases, participants were told that they were randomly assigned to one of 5 current media topics when in fact, the topic (fictitious

chemical, tryptostine) remained identical for all participants (van der Linden et al., 2017). They were then asked to complete a pre-test, measuring attitudes towards said fictitious chemical (Tryptostine), their attitude certainty, and their perceived ability to resist persuasion. Participants were then randomly assigned to one of two conditions (inoculation, control). The inoculation condition consisted of a treatment message which included a warning about imminent persuasion attempts and pre-emptively debunked (i.e., prebunked) common manipulative strategies. The control condition received a neutral article about the fictitious chemical following the practices of previous research on information transmission through inoculation (e.g., Ivanov et al., 2016).

Subsequently, all participants were exposed to an attack message aiming to spread misinformation about the issue topic through the use of manipulative. To summarise, one inoculation message (411 words), one attack message (390), and one control message (370 words) were devised for this pre-registered study<sup>31</sup>. The misinformation message has been produced by following the practices of previous research on information transmission (Moussaïd et al., 2015) and, for ecological validity, was created by basing it on misinformation about a real chemical that went viral in the past<sup>1</sup> and arguments around reputable sources on the use and safety of specific chemicals in consumer goods (example sentence: "Most people are under the impression that chemicals are vigorously tested before they are allowed on the market but this is simply not true. In fact, Tryptostine, the main ingredient used in most antibacterial products was suspiciously never vetted by the U.S. Food and Drug Administration (FDA) at all.").

Additionally, while both the attack message and control message focused on the fictitious chemical, the inoculation treatment followed previous research (van der Linden & Roozenbeek, 2019; Basol et al., 2020; 2022) and was composed to inoculate against techniques used in the spread of misinformation (emotional language, fake experts, conspiracy theories). As discussed in Chapter 4, these three strategies have been reported as pivotal to the spread of health misinformation during the pandemic. Specifically, research suggests that the use of moral-emotional language enhances the virality of social media content (Acerbi, 2019; Berriche & Altay, 2020; Brady et al., 2017), that manipulative health content relies on techniques using impersonation and fake experts (Cook et al., 2017; Roozenbeek & van der Linden, 2019), and that conspiracy theories are heavily featured around

<sup>&</sup>lt;sup>31</sup> A pilot study (N=100) was conducted to assess the persuasiveness and perceived credibility of each treatment message.

health misinformation (Ball & Maxmen, 2020; Lazić & Žeželj, 2021). Thus, considering the study which tested the efficacy of inoculating against these strategies in the context of the pandemic (Basol et al., 20210), the current chapter applies these strategies to the context of a fictitious, yet a seemingly dangerous chemical. Importantly, these strategies corresponded with the tactics used in the attack message, making the inoculation treatment a generalised rather than issue-specific one (Compton et al., 2020; Roozenbeek & van der Linden, 2019; Basol et al., 2021). Put differently, the inoculation treatment begins with the traditional element of explicit threat and then pre-emptively refutes the same manipulation strategies but applied to a different topic (e.g., "Most of us are confident in our ability to detect and resist misinformation and we wouldn't think about sharing it. Whether you are aware of it or not, there is a high likelihood that in the past you have fallen for and perhaps even shared misinformation with others. [...] One way to make falsehoods go viral is by using **emotional language**. Emotional content is not necessarily "fake" or "real" but rather deliberately plays into people's basic emotions such as fear, anger, and empathy [e.g., "Shocking results on the side-effects of cotton ear buds in this recently leaked report!"]).

Doing so adds to the ongoing scientific exploration of the blanket of protection, where individuals are inoculated against unmentioned arguments within the same issue topic or the phenomenon of cross-protection where inoculation-conferred resistance extends to related yet untreated topics (Traberg et al., 2022). After exposure to the misinformation message, all participants completed the post-test (identical to pre-test) as well as a set of additional measures (counterarguing, motivational/apprehensive threat). Importantly, participants were asked to complete the first measure on PIT (*intentional pass-along message*) – here, participants were required to report whether and what they would share online with others about the treatment message (for all measures, see below). Finally, participants received a generalised debrief form and were reminded that they would be reinvited to the pre-registered follow-up study a week later (for full study flow, see Figure 25). Phases 1 and 2 were pre-registered (https://aspredicted.org/a8y52.pdf) and the studies presented in this chapter were approved by the Cambridge Ethics Committee (PRE.2021.069).

# Sample

Participants were recruited via *Prolific Academic* (Peer et al., 2017). Participants who completed the full study (including the follow-up study) received a total of £1.45 in compensation. Participants who

submitted low-effort responses or failed to pass the attention check<sup>32</sup> were excluded from the analysis. Taking the post-inoculation talk effect reported in previous research (Ivanov et al., 2016), G\*power with alpha = .05, f = .25, and power of .95 with 2 experimental conditions and repeated measures, the projected sample size needed is approximately N = 158. In total, 401 participants (200 per condition) were recruited and 357 completed the follow-up study (11.6 % attrition). Of the sample, 50.6% identified as female (48.6% male, 0.9% other); 90.1% indicated being between 21 and 37 years of age, and 46% reported having a bachelor's degree. Lastly, the sample skewed politically left (M = 3.33, SD= 1.4).

<sup>&</sup>lt;sup>32</sup> This was a simple multiple-choice question asking participants to indicate the topic of the text (i.e., Genetically Modified Food, the chemical tryptostine, or techniques to spread misinformation.



\*Sample A was blacklisted on Prolific Academic.

Figure 25: Study flowchart for all three phases.

Pre-post test Attitude change Attitude confidence Perceived resistance

### AM

Counterarguing Motivational threat Traditional threat Passing on message

#### Follow-up

Post-test Counterarguing Motivational threat Traditional threat PIT Frequency of discussions N of links

119

## Measures

### Attitudes towards target topic

Following prior inoculation research (e.g., Ivanov et al., 2013), participants rated the target topic (tryptostine) using seven semantic differential items (negative/positive, bad/good, dislike/like, desirable/undesirable, unfavourable/ favourable, unacceptable/acceptable, and wrong/right). All ratings were measured through 7-point scales with larger numbers reflecting a more positive attitude towards target topics. The attitude scale was found to be highly reliable (7 items;  $\alpha$ = 0.89).

### Attitude change

This measure was obtained by subtracting the post-test attitude measure from the initial pre-treatment attitudes (pre-test). This resulted in a treatment-induced attitude change index value. An index of zero indicates no attitude change from phase 1 to the follow-up. Positive index values indicated a strengthening of initial attitudes, while negative values signalled the weakening of initial attitudes, respectively. To control for initial (Phase 1) attitudes (M = 3.92; SD = 0.83), and capture the inoculation moderated change, the final (Phase 2) postattack attitudes (M = 3.21; SD = 1.13) were subtracted from the initial (Phase 1) pre-inoculation attitudes, with the resulting product comprising an inoculation-moderated attitude change index value. Thus, an index value of zero indicates no attitude change from Phase 1 to 3. Positive index values indicate a strengthening of initial attitudes, whereas negative values signal the weakening of initial attitudes (M = -0.71; SD = 1.24).

### Attitude strength

Participants responded to six self-report measures of attitude strength frequently used in attitude strength research (e.g., Krosnick et al., 1993; Brannon et al., 2007). Participants' attitude strength was tested after treatment exposure and in the follow-up and tapped into attitude certainty (e.g. "How sure are you that your opinions about tryptostine are right?"), personal importance ("How important to you personally is the issue of X?"), ego involvement (e.g. "How representative of your values is your attitude toward X?"), issue knowledge ("How knowledgeable are you about X?"). Considering the ongoing debate concerning the appropriateness of attitude extremity as an indicator of attitude (Eagly & Chaiken, 1997), this research followed recent attitude strength research (Brannon et al., 2007) and did not include attitude extremity. The attitude strength scale produced good reliability ( $\alpha = .72$ ).

#### Attitude confidence

This measure contained a single question: "How certain are you of your opinion toward the use of chemicals such as triclosan in consumer goods?" This item was adapted from previous research (Tormala & Petty, 2004) and provided responses on a 1-7 scale (1= not at all certain; 7= very certain).

#### Motivational threat

This measure was implemented from previous studies (Banas & Richards, 2017; Richards & Banas, 2018) and asked participants to consider the treatment message (inoculation, misinformation) and state their agreement with four statements (e.g., "The message motivated me to resist reacting negatively to information about [issue topic].") on a 7-point scale (1=strongly disagree; 7= strongly agree). The four items demonstrated acceptable internal consistency,  $\alpha = .69$ . As argued by Hair and colleagues (2010), values as low as 0.6 are acceptable for exploratory research.

#### Apprehensive threat

This measure consists of six items on a seven-point semantical differential scale (Compton & Pfau, 2005; Parker et al., 2012b). The item traditionally states: "The next set of items are designed to help us to understand how you feel about the idea expressed at the beginning of the message you just read that, despite your opinion on this issue, there is a possibility you may come into contact with arguments contrary to your position that are so persuasive that they may cause you to rethink your position. I find this possibility..." and is followed by bipolar adjectives (non-threatening/threatening, not harmful/harmful, not risky/risky, not dangerous/dangerous, calm/anxious, and not scary/scary). These items also demonstrated excellent internal consistency,  $\alpha = .93$ .

### Counterarguing

Counterarguing in response to the attack messages was evaluated with a single-item measure (Ivanov et al., 2016). After reading the attack message targeting each respective topic, participants were asked to report the degree to which they counterargued against the attack message on a 7-point scale with the anchors "I accepted a lot of the arguments offered" (1) and "I thought of a lot of arguments against [the arguments offered]" (7). A single-item measure was used in this study to mitigate participant fatigue, given that participants were asked to report on four topics. This measure has been shown to be correlated with the results of open-ended counterarguing measures used in prior inoculation research (Miller et al., 2013).

#### Perceived resistance

A single-item measure was used to determine participants' perceived ability to resist information against the fictitious chemical ("Overall, how difficult was it for you to come up with arguments **against** the message about the **dangers of tryptostine**?"), which they answered on a 7-point scale with the anchors "Not difficult at all" (1), "Neither difficult nor easy" (4), and "Very difficult" (7).

#### Sharing intentions

A single-item measure (M= 4.0; SD= 1.91) was used to examine participants' willingness to talk to others about what they've learned about the chemical ("You've read one or more articles today. Please indicate your willingness to share what you have learned in the text(s) with others (i.e., by telling them

about it in real life)"), which they indicated on a 7-point scale with the anchors "Not at all willing" (1) to "Very willing" (7).

#### Intentional pass-along messages

All participants were asked to complete a text entry task in which they were prompted to compose a tweet-sized message about the study's content to be passed on with the next round of participants ("Imagine you decided to tweet about what you have learned from the messages you've just read – within 140 characters, what would you like to share with others?"). This measure was limited to 140 characters to mimic the length of a Tweet. Phases 1 and 2 introduce a variety of PIT measures of which all can be reviewed in the Table 1.

# Content analyses and coding procedures

As per pre-registration, the intentional pass-along messages collected in Phase 1 served to develop a codebook that identified themes and reoccurring terminologies which would allow for examinations of the content of participants' messages (see Table 1, for examples). To do so, a coding scheme of a series of categories was developed to be mutually exhaustive and exclusive (Krippendorff, 1980). Accordingly, the intentional pass-along messages were judged on the basis of whether they repeated or added to the misinformation in the attack message (1=yes; 0= no; M=0.43; SD=0.47), whether they displayed basic resistance (1=yes; 0=no; M=0.51; SD=0.49), and whether they demonstrated enhanced resistance (2= inoculation-congruent; 1= basic resistance; 0= no resistance; M=0.59; SD=0.63). Table 5.1. presents scoring examples. To account for the complexity of messages, where, to give an example, a pass-along message can contain both inoculation-congruent material (i.e., prebunking, forewarning) while also still repeating misinformation, the 'degrees of resistance' index was designed (M=0.01; SD=1.44). Here, rather than a dichotomous conceptualisation of spreading misinformation vs truth, the pass-along messages were scored on a continuous scale ranging from conspiratorial thinking (-3), spreading misinformation (-2), repeating misinformation (-1), sharing nothing (0), general resistance (1), enhanced resistance (2), and advocative resistance (3). See Table X for breakdown and scoring procedures of these categories.

# Coders, coding procedures

After the codebook development, an electronic file containing all messages composed by participants at different time points was developed and, alongside the above-described coding scheme, shared with two researchers. Importantly, the document containing the pass-along messages omitted the treatment condition or any other information that could influence the coding process. Prior to any coding, both researchers went through 5 responses together to establish an understanding of the coding procedure.

# Intercoder reliability

Intercoder reliability was calculated using Cohen's  $\kappa$  Kappa (1960). Both researchers independently completed the coding for all measures of PIT in this chapter (intentional pass-along message, actual PIT, and intentional pass-along message of vicariously inoculated sample). Intercoder reliability assessments demonstrated substantial to almost perfect agreement, ranging from  $\kappa = 0.69$  to  $\kappa = 0.92$  (see supplements for all reliability measures).

Accordingly, the following pre-registered hypotheses were tested:

H<sub>1</sub>: Inoculated individuals will experience greater resistance than the control condition.

**H<sub>2</sub>:** Compared to the control condition, inoculated participants will report no or positive attitude change.

 $H_3$ : Inoculated individuals will be significantly more confident in their held attitude than the control group.

 $H_4$ : Inoculated individuals will score significantly higher in motivational and apprehensive threat than the control condition.

 $H_5$ : Higher levels of motivational threat will predict perceived resistance to messaging attacking one's beliefs.

H<sub>6</sub>: Inoculation-induced motivational threat will be positively associated with an increase in PIT.

 $H_7$ : After a week, participants in the inoculation condition will report significantly more post-treatment talk (both in terms of conversational partners and frequency of conversations) than participants in the control conditions.

H<sub>8</sub>: Among inoculated individuals, post-inoculation talk will be positively associated with resistance.

**H**<sub>9</sub>: A week later, inoculation individuals who engaged in PIT will score significantly higher in attitude strength, confidence, and perceived resistance than the control condition.

 $H_{10}$ : In their pass-along messages, inoculated participants will share less content from the misinformation message than individuals in the control condition.

Table 1: Summary of all PIT measures with scoring examples.

| +                                  |  |   |                        |  |  |  |  |  |  |
|------------------------------------|--|---|------------------------|--|--|--|--|--|--|
| N                                  | Ieasure                                | Description   | Testing-<br>point      | Scoring examples   |  |  |  |  |  |
| Freque<br>inocul<br>conver         | ency of<br>lation<br>rsations          | Since last week, <b>how many conversations</b> have you had about<br>common <i>manipulation techniques</i> used to spread misinformation?<br>(Note: this can be different from how many people you talked to,<br>e.g., several conversations with one individual) | <i>T</i> 2             | In the first four measures listed here, participants indicate the number of conversations/conversation partners on a drop-down multiple-choice question (numbers ranging from 0 to 20+). See Phase 2, "Traditional PIT" for more details.  |  |  |  |  |  |
| Numb<br>conver<br>partne<br>(inocu | er of<br>rsation<br>ers<br>llation)    | Since last week, <b>how many people</b> did you tell about the common <i>manipulation strategies</i> used to spread misinformation?   | <i>T</i> 2             |  |  |  |  |  |  |
| Freque<br>misinf<br>conve          | ency of<br>formation<br>rsations       | Since last week, <b>how many conversations</b> have you had about the dangers of <b>tryptostine</b> ? (Note: this can be different from how many people you talked to, e.g., several conversations with one individual)   | <i>T</i> 2             |  |  |  |  |  |  |
| Numb<br>conver<br>partne<br>(misin | er of<br>rsation<br>ers<br>aformation) | Since last week, how many people did you tell about the dangers of <i>tryptostine</i> ?   | <i>T</i> 2             |  |  |  |  |  |  |
| Intentialong                       | ional pass-<br>message                 | Imagine you decided to tweet about what you have learned from the messages you've just read – within 140 characters, what would you like to share with others?  | Т1, Т3                 | These text box entries were independently coded by two researchers to develop<br>the measures "spreading inoculation", "spreading resistance", "spreading<br>misinformation" as well as the "degrees of resistance" index.   |  |  |  |  |  |
| Spread                             | ding<br>lation                         | (0=not sharing anything, 1= basic resistance (see below), 2=<br>resistance + inoculation)   | <i>T</i> 1, <i>T</i> 2 | <ul> <li>Googling chemical + warning that chemical is fake/made-up</li> <li>Traditional components of inoculation, e.g., element of forewarning ("be aware!" "They will tell you that" and pre-emptive refutations (explicitly prebunking content/manipulation techniques)</li> <li>Examples: "The chemical tryptostine is misinformation", "misinformation is a threat, always check the sources and verify statements", "People who are against tryptostine use inflammatory and emotionally charged language to get you on their side"</li> </ul> |  |  |  |  |  |

| Spreading basic   | (0= no resistance, 1= any form of resistance)                         | $T_{1, T_{2}}$ | - Any form of resistance that is not inoculation-specific                      |
|-------------------|---|----------------|--|
| resistance        |   |                | <ul> <li>General commentary on ads, algorithms, online manipulation</li> </ul> |
|                   |   |                | - Reassurance-seeking behaviour: Requesting more information about             |
|                   |   |                | googling chemical  |
|                   |   |                | - Not wanting to share things one isn't sure about                             |
|                   |   |                | - Examples: "Not sure", "supposedly", "does anyone know more", "need           |
|                   |   |                | look into this further", "didn't share because I don't know enough about       |
|                   |   |                | "google says", "I trust experts", "wouldn't believe just everythin             |
|                   |   |                | read", "always check other sources"  |
| Spreading         | (0 = no misinformation, 1 = sharing misinformation)                   | T1, T2         | - Simple repetition of claims made in attack message (e.g., that chemica       |
| misinformation    |   |                | dangerous, needs to be removed, has terrible side effects)                     |
|                   |   |                | <ul> <li>made up information about topic issue</li> </ul>                      |
|                   |   |                | - conspiratorial content (e.g., blaming gov, FDA, 'higher powers', "we         |
|                   |   |                | their guinea pigs")  |
|                   |   |                | <ul> <li>spreading doubt, urging others to stop using the chemical</li> </ul>  |
|                   |   |                | <ul> <li>asking for the fictitious chemical to be banned</li> </ul>            |
|                   |   |                | - examples: "be aware", "you won't believe", "out to kill u                    |
|                   |   |                | "dangerous/harmful", "no one is talking about it", "not tested", "it sho       |
|                   |   |                | be banned"   |
| Resistance Index  | (-3 = conspiratorial thinking, -2 = spreading Misinfo, -1 = repeating | $T_{1}, T_{2}$ | Exemplary scoring  |
|                   | Misinfo, 0= sharing nothing, 1= general resistance, 2=enhanced        |                | - "Tryptostine is out to kill us all and the government is funding it" (-3)    |
|                   | resistance, 3= resistance + advocative inoculation)                   |                | - "I recently read an article about the dangers caused by this chemical an     |
|                   |   |                | AM SHOCKED! How can governments allow that!!" (-2)                             |
|                   |   |                | - Tryptostine was never vetted by the FDA. Tryptostine can be found in ha      |
|                   |   |                | sanitisers, soaps and body washes, toothpastes, cosmetics and more (-1)        |
|                   |   |                | - There are several opinions about tryptostine, a food additive that may have  |
|                   |   |                | impact on human life. (0)  |
|                   |   |                | - "What's the deal with this tryptostine debate? Can anyone comment? I have    |
|                   |   |                | read so much stuff about it i have no idea where to stand." (1)                |
|                   |   |                | - "Some people have been spreading information and misinformation about        |
|                   |   |                | chemical that does not exist. Did you fall for it?" (2)                        |
|                   |   |                | - "Fake news uses fake authority, emotions and conspiracies, so if you rea     |
|                   |   |                | article about a chemical that uses all that, maybe don't trust it" (3)         |
| Actual pass-along | In one sentence, what did you tell others about common                | Т2,            | - Scoring actual pass-along message on above mentioned measures                |
| message           | manipulation strategies in the last week?                             |                | (spreading basic resistance/ misinformation/ inoculation and resistance)       |
|                   |   |                | and index.   |

# Results

Firstly, a series of preregistered analyses for two main outcome measures (resistance, PIT) will be presented, focusing primarily on the *difference* for each outcome, before and after the intervention between conditions (i.e., hypotheses  $H_1$ - $H_6$ )<sup>33</sup>. Additionally, some initial insights into *intentional* passalong messages will be reviewed to gauge what inoculated individuals *would* share with others online.

## Resistance

A one-way between-subjects ANOVA showed a significant effect of condition (Inoculation, Control) on the perceived resistance against misinformation about the chemical, F(1,396)=9.78, p=0.002, ,  $\eta^2 = 0.024$ . Results suggest that inoculated individuals are perceive their resistance as significantly higher than the control condition (M<sub>control</sub>= 3.43, M<sub>inoc</sub>=3.91, M<sub>diff</sub> = -0.479, 95% CI (-.78, -.17), d= -0.31). To see whether this observed effect changed over time, a repeated measures one-way ANOVA was conducted with condition (inoculation, control) as the between-subject factor and time (T<sub>1</sub>, T<sub>2</sub>) as the within-subject factor. Doing so illustrates a significant effect of time x condition on the perceived resistance against misinformation, F(1,352)=6.37, p=0.01,  $\eta^2 = 0.01$ . These findings support H<sub>1</sub>: showing that inoculated individuals experienced greater resistance than the control condition.

<sup>&</sup>lt;sup>33</sup> Robustness checks for these results, including how they differ across covariates (age, gender, education, etc.) can be found in the Supplementary Analyses section on the OSF page.



Figure 26: Violin plot with jitter of perceived resistance scores between conditions.

# Attitude change

A one-way between-subjects ANOVA does not show a significant effect of condition (Inoculation, Control) on the pre-post intervention difference of attitudes towards the fictitious chemical, F(1,375) = 0.49, p=0.48, d=.07,  $\eta^2 = 0.001$ . To test whether attitudes towards the fictitious chemical changed over time, a repeated one-way ANOVA was conducted with condition (inoculation, control) as the between-subject factor, and time (pre-treatment, post-, follow-up) as the within-subject factor. Mauchly's Test of Sphericity indicated that the assumption of sphericity was violated,  $\chi^2(2) = 0.85$ , p < .001 and therefore, a Greenhouse-Geisser correction was used. At the within-subject level, a significant effect of time is evident, F(1.7, 599.3) = 78.56, p = 0.001,  $\eta^2 = 0.08$ . Additionally, there is no significant interaction of time x condition on the issue topic, F(1.7, 599.3) = 0.55, p = 0.54,  $\eta^2 = 0.001$ . Thus, no support is found for **H**<sub>2</sub>: there is no significant difference in attitude change between conditions.

# Attitude certainty

To test whether the certainty with which attitudes towards the fictitious chemical are held changed over time, a repeated one-way ANOVA was conducted. Specifically, with condition (inoculation, control) as the between-subject factor, and time (post-treatment [T<sub>2</sub>], follow-up [T<sub>3</sub>]) as the within-subject factor. The results demonstrate no significant interaction between time x condition on attitude certainty, F(1,352)=0.06, p=0.79,  $\eta^2=0.001$ . Consequently, there is no support for H<sub>3</sub>: inoculated individuals do not report significantly higher attitude confidence than the control group.

# Apprehensive threat

To assess whether individuals' perception of (traditional) threat differed between experimental conditions, a one-way ANOVA analysis was conducted. The results demonstrate a significant main effect of condition on apprehensive threat, F(1, 396) = 40.9, p=0.001, d=-0.64. Tukey's post-hoc analyses show that inoculated individuals reported significantly higher threat levels than the control condition ( $M_{\text{control}}$  = 3.05,  $M_{\text{inoc}}$  = 3.94,  $M_{\text{diff}}$  = -0.89,  $p_{\text{tukey}}$  = 0.001, 95% CI [-1.16, -0.61] d = -0.64). Additionally, a one-way repeated measures ANOVA was run to assess whether perceived threat of having one's attitudes attacked changed over time. Hence, condition (inoculation, control) was used as the between-subject factor and time (post-treatment  $[T_2]$ , follow-up  $[T_3]$ ) as the within-subject factor. The results show a significant interaction between time x condition on perceived apprehensive threat, F(1,352) = 39.4, p = 0.001,  $\eta^2 = 0.03$ . Additional Tukey's post-hoc analyses show that the significant difference occurs on the within-subject level. More specially, while inoculated individuals report significantly lower threat levels compared to immediately after treatment exposure  $(T_2)$ , participants in the control condition report significantly higher apprehensive threat levels in the follow-up ( $M_{\text{inoc},T2} = 3.9$ ,  $M_{\text{inoc},T3} = 3.35$ ,  $M_{\text{diff},\text{inoc}} = 0.55$ , p = 0.001, 95% CI (-.76, -.34); *M*<sub>controlT2</sub> = 3.03, *M*<sub>control,T3</sub> = 3.52, *M*<sub>diff, control</sub> =- 0.49, *p* = 0.001, 95% CI (0.23, 0.75).



Figure 27: Line graph of motivational and apprehensive threat (T1, T2) between conditions.

## Motivational threat

To assess whether there is a main effect of condition (inoculation, control) on motivational threat, a one-way ANOVA was used. Results suggest a significant main effect of condition on participants' self-reported perceptions of motivational threat, F(1,395)=4.56, p=0.03,  $\eta^2=0.01$ . Additional Tukey's post-hoc comparisons show that the inoculation condition reports significantly higher on motivational threat than the control condition (M<sub>control</sub>= 4.46, M<sub>inoc</sub>=4.69, M<sub>diff</sub> = -0.22, p<sub>tukey</sub> = 0.03,

95% CI [-0.41, -0.01] d= -0.21). Furthermore, to assess whether the effect of condition on motivational threat changed over time a one-way repeated measures ANOVA was used with condition (inoculation, control) as the between-subject factor, and time (T<sub>2</sub>, T<sub>3</sub>), as the within-subject factor. Results demonstrate no significant interaction between time x condition on motivational threat, F(1,352)=0.41, p=0.52, ,  $\eta^2 = 0.001$ . Consequently, these results provide support for **H**<sub>4</sub>: inoculated individuals scored significantly higher on motivational and apprehensive threat than in the control condition. Lastly, linear regressions show that threat predicts perceived resistance against misinformation about the issue topic,  $R^2 = .04$ , F(3, 394) = 5.78,  $p = .001^{34}$ . Examining the individual predictors indicate that condition (Beta=0.54, *t* (0.16) =3.39, *p*= 0.001), motivational threat (Beta=.16, *t* (.07) =2.21, *p*= 0.02), and apprehensive threat (Beta= -.11, *t* (0.05)=-2,01, *p*= 0.04), were significant predictors in the model. Thus, these findings provide support for **H**<sub>5</sub>: heightened motivational threat predicted perceived resistance against misinformation about the issue topic.

Moreover, a series of analyses were conducted to examine the role of motivational threat in post-inoculation talk. Post-inoculation talk was measured at several time points  $(T_1, T_2)$  and in multiple ways (i.e., sharing intention, frequency and number of conversations, text box entries). A word cloud plot of the frequency of words occurring in the text box entries can be found below. Thus, examining the effects of motivational threat on PIT will involve different attempts to operationalise postinoculation talk. During the first phase of the study, post-inoculation talk was measured in two ways (sharing intention, and an intentional pass-along message). A linear regression shows that motivational threat significantly predicts individuals' intentions to talk to others about the treatment message  $(T_1)$ ,  $R^2 = .057, F(3, 394) = 7.98, p = .001$ . Additionally, at the individual level, motivational threat is a significant predictor in the model (Beta=0.34, t (0.12) =3.74, p= 0.006). Indeed, a one-way ANOVA demonstrates a significant main effect of condition on sharing intentions, F(1,396) = 15.1, p = 0.001,  $\eta^2 = 0.038$ . Additionally, Tukey's post-hoc comparisons suggest that the control condition scores significantly higher on intentions to share through talk than the inoculation condition (M<sub>control</sub>= 4.36,  $M_{inoc}$ =3.63,  $M_{diff}$  = 0.73,  $p_{tukey}$  = 0.001, 95% CI [0.19, 0.58] d= 0.39). When conducting an ANCOVA with sharing intentions as the dependent variable, condition as the independent variable, and motivational threat as a covariate, the results suggests that there is no significant interaction between

 $<sup>^{34}</sup>$  Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (Motivational threat, Tolerance = .94, VIF = 1.05; Apprehensive threat, Tolerance = .86, VIF = 1.15)

threat x condition, F(1,394) = 1.42, p = 0.23,  $\eta^2 = 0.004$ . Consequently, these results fail to find initial support for **H**<sub>6</sub>: Inoculation-induced motivational threat is not positively associated with an increase in post-inoculation talk conceptualised through sharing intentions.

### The spread of inoculation vs. misinformation

First, a word cloud plot (Figure 28) was created to provide a more nuanced understanding of the words that occurred most when individuals reflected on what they would share with others (the bigger the font size, the more frequent use of that word). Additionally, to assess whether the intentional passalong messages composed after treatment and attack message exposure differed significantly in the content they were spreading (inoculation vs. misinformation), a series of exploratory ANOVA's were conducted on the combined scorings of the messages (average of scores from both independent researchers). First, at the general resistance level (i.e., any messages that could have been produced without exposure to specific inoculation message), results show a marginally significant main effect of condition on post-treatment talk that spreads resistance, F(1,394) = 3.95, p = 0.048,  $\eta^2 = 0.01$ . Additionally, Tukey's post-hoc comparisons suggest that the messages by inoculation individuals score spread significantly more basic resistance than the control condition (M<sub>control</sub>= 0.46, M<sub>inoc</sub>=0.56, M<sub>diff</sub> = -0.09,  $p_{tukey} = 0.048$ , 95% CI [0-.39, -0.002] d = 0.2).

When assessing whether there is a difference in whether messages intend to pass on inoculation-congruent content (i.e., forewarning, pre-emptive refuting, list of manipulation techniques), the results demonstrate a significant main effect of condition on the spread of inoculation, F(1,394) = 16.0, p = 0.001,  $\eta^2 = 0.039$ . Additionally, Tukey's post-hoc comparisons suggest that the inoculation condition scores significantly higher on intentions to share through talk than the inoculation condition (M<sub>control</sub>= 0.467, M<sub>inoc</sub>=0.72, M<sub>diff</sub> = -0.25,  $p_{tukey} = 0.001$ , 95% CI [0-.6, -0.2] d = 0.4). On the other hand, no significant main effect of condition on the spread of misinformation about the fictitious chemical in intentional pass-along messages is evident, F(1,394) = 1.56, p = 0.21,  $\eta^2 = 0.004$ .

trust something conspiracy tryptostin always effects say called companies maybe using life anyone bad sources government scientific tested tweet even learn share misinformation serious topic human want anything around research everyone dangerous careful articles substance everything studies every believe used media reading seems without tryptosine find kno compound cause safety harmful 11Se aware health just food article heard information thing good others dangers trypostine peop need sure safe fake made news false keep one really cts fda proc nothing things lot opinion important chemicals body true bodies approved daily everyday certain product humans consume cosmetics

Figure 28: Word cloud plot for intentional pass-along messages composed at T1.

## Discussion - Phase 1

Phase 1 aimed to investigate  $H_1 - H_6$  and finds no significant within-subject differences (pre-post) in attitude change, attitude certainty, nor attitude strength. However, inoculated individuals did perceive their resistance against misinformation about the issue topic to be significantly higher than the control condition. In addition, inoculated individuals also experienced significantly higher levels of apprehensive and motivational threat. Higher levels of the latter predicted perceived resistance as well as an increase in intended post-inoculation talk. Lastly, examining the content of the pass-along messages composed by each participant suggests that inoculated participants did spread significantly more content on resisting misinformation in general (e.g., fact-checking online content) and inoculation-specific content (i.e., forewarning and pre-bunking strategies) than the control condition. However, when asked what they would pass on to others, no significant difference in the repetition or spread of misinformation about the issue topics was evident between the conditions.

# The origins of post-inoculation talk

These findings are partially in support of the pre-registered hypotheses and the limited existing research on post-inoculation talk. Traditionally, inoculation scholarship measures attitudinal resistance through examining changes in self-reported attitudes towards the issue topic, perceived resistance (Ivanov et al., 2012), and perceived ability to counterargue (Banas & Miller, 2013). Since conditions did not differ in their attitude changes, the initial assumption would perhaps be that the inoculation treatment failed to confer resistance against the issue topic. On top of that, content analyses of the *intended* pass-along messages composed by both conditions after exposure to the misinformation message (i.e., attack message) suggest that there was no difference in how often these

messages were repeated or built on the misinformation about the issue topic. In other words, while on the one hand, the absence of attitude change would traditionally be conceptualised as attitudinal resistance, the entire sample still included parts of the attack message in their intentional pass-along messages. However, while no noticeable difference in attitude change seems to occur between conditions, the results do show that inoculated individuals *perceive* their ability to resist and refute manipulative arguments about the issue topic to be significantly higher than the control condition. Similarly, pass-along messages by inoculated individuals spread significantly more basic resistance (e.g., asking for further evidence, googling the chemical) and inoculation-congruent content (e.g., forewarning others about manipulation techniques).

There are several potential explanations for this. To begin with, given the fictitious nature of the issue topic, participants did not have any pre-existing attitudes at the beginning of the study. Consequently, issue involvement, often regarded as an essential factor in the inoculation process (Pfau et al., 2005), was not incorporated in this study. Considering that the present studies in this chapter investigate the propagation of inoculation when pitted against misinformation, the risks of having participants who were not exposed to inoculation treatments be exposed to, fall for, and consequently, spread misinformation about true topics were too high. Future research should investigate whether prioritising ecological validity and the ethical commitment to not contribute further to the spread of (even fictitious) misinformation undermines the efficacy of an inoculation treatment. For this study, these were the only parameters possible given the ethics committee's concerns about spreading further misinformation during a global health crisis.

### Rethinking resistance

On the other hand, a fictitious topic allowed to exclude memory confounds (Hassan & Barber, 2021) to influence attitude change and attitude certainty as well as whether and what participants decide to pass on to others. After all, in today's fast-paced information infrastructure, it is crucial for inoculation treatments to keep up with exposure to manipulation techniques, fabricated content, and conspiracy theories (Bleakley, 2021; Douglas et al., 2019; Rodny-Gumede, 2018). Thus, if inoculation treatments are to have a chance against online misinformation, examining PIT without issue involvement offers a more ecologically valid and realistic setting to test the efficacy of inoculation against the rise and spread of misinformation about various (fictitious) topics. Lastly, it could be that immediate administration of the attack message shortly after treatment exposure did not provide sufficient delay for an opinion on a previously unknown topic to be made, let alone changed. Though a meta-analysis (Banas & Rains, 2010) finds no support for the effects of treatment delays on attitudinal resistance, the particular context of a fictitious issue topic in this study might be affected differently by (the lack of) delays. Thus, instead of questioning whether the inoculation treatment was successful in conferring resistance against misinformation about the fictitious chemical, the lack of attitude change in the control condition suggests that single exposure to information about it might not suffice to change

one's opinion – especially when no opinion existed up until the start of the study altogether. However, since one essential focus of this chapter was to emulate and test the efficacy of inoculation treatments in a social-media environment, it can be argued that keeping the delay between treatment and attack message exposure small is consistent with the way in which one might encounter pre-bunking and misinforming content in their newsfeed. Furthermore, since inoculated participants reported significantly higher levels of *perceived* resistance against misinformation about the issue topic and demonstrated a significant spread of resistance and inoculation in their intended pass-along messages, one might also question whether the current conceptualisation of attitudinal resistance lacks breadth.

# What is PIT all about?

Consequently, content-level analyses on intended post-treatment talk emphasise the need for more nuanced snapshots of attitudinal resistance. These entries allow an exploration of the extent to which PIT is in alignment or in contradiction with the inoculation treatment. Indeed, the word cloud plot provides an initial insight into the content of these intended pass-along messages. On the other hand, the independent scorings of the messages on the 'degrees of resistance' scale point towards the need for clearer categorisations when studying the propagation of misinformation. Or rather, scales that make room for the more ambiguous engagement with the inoculation treatment and the misinformation about the issue topic. For instance, many participants' PIT messages repeated parts of the misinformation message yet were followed by expressions of uncertainty and information-seeking attempts (e.g., "Does anyone know more about this?", "Could someone help me understand?"). Though the current study design does not allow to test this, it is, therefore, possible that the inoculation condition spreads misinformation as a consequence of threat-driven reassurance-seeking via talk (Compton & Pfau, 2009). This is consistent with the previously noted significance in the inoculation process, suggesting that motivational threat is not only crucial to the development of "cognitive antibodies" and attitudinal resistance but possibly plays a substantial role in the strengthening and spreading of inoculation as well. It could even be argued that this is reflected in the absence of changes occurring in attitude certainty and attitude strength, as noted in previous chapters. Thus, this further emphasises the need for measurements and indexes that account for the sometimes unclear if not contradictory nature of the spread of misinformation.

# Challenging the boundary conditions

To take it further, one could argue that the full breadth of attitudinal resistance, too, remains unexamined, if not misunderstood – especially in relation to the more novel applications of inoculation treatments against online misinformation. Inoculation scholarship has experienced a re-emerged interest in the last few decades and as a result, there is a large body of research (for a comprehensive review, see Ivanov et al., 2020; Compton et al., 2021) that has demonstrated the effectiveness of inoculation treatments in contexts ranging from legalising animal testing (Nabi, 2003) to contested issues such as climate change and vaccines (Jolley & Douglas, 2017; van der Linden et al., 2017).

However, the current methodological parameters prevent the scholarship and policy-makers from making full use of the potential that inoculation holds. While the gamified, generalised, and therapeutic inoculation treatments discussed in the previous chapters offer crucial theoretical and practical advancements, it remains unclear as to how inoculation theory can fully adapt to and optimise within the context of today's information infrastructure. Therefore, by using a treatment message that is not issue-congruent with the attack message but instead pre-bunked the same manipulation techniques in untreated health topics (e.g., side-effects of cotton ear buds, 5G radiation), the present study tests the ability of inoculation messages conferring cross-protection (Ivanov et al., 2012; 2016). The findings suggest that the treatment message was effective in conferring resistance against the untreated issue topic of a fictitious chemical. In fact, observational analyses demonstrate how many pass-along messages by inoculated individuals specifically warned against the fictitious chemical or directly applied the learned manipulation techniques to the context of the issue topic (e.g., "Fake news uses fake authority, emotions and conspiracies, so if you read 1 article about a chemical that uses all that, maybe don't trust it" and "People who are against Tryptosine use inflammatory and emotionally charged language to try and get you on their side."). Thus, although such insights are anecdotal at this stage, they offer initial insights into the ability to confer "cross-protection" against previously unmentioned and unrelated topics.

### Conclusion

Phase 1 has provided initial steps toward establishing a more nuanced understanding of postinoculation talk – that is, one that goes beyond the traditionally constrained snapshot of frequency and number of conversations. The results suggest that inoculated individuals feel more motivationally threatened and report a greater perceived ability to resist and refute manipulative content about the issue topic than the control condition. Additionally, when asked what they would hypothetically like to pass on to others, intentional pass-along messages composed by inoculated individuals shared more general resistance and resistance-congruent content. However, since the intentional pass-along message was based on a hypothetical scenario, it is unclear whether inoculated individuals will actually engage in post-inoculation talk in the interim between Phases 1 and 2. Additionally, on top of not knowing whether intended talk did indeed occur, more insights into the content of their actual talk is needed. In short, if inoculated participants did voluntarily engage in post-inoculation talk during the interim, how many conversations did they have, how many different people did they voluntarily speak to, and more importantly, what did they *actually* talk about? By distinguishing between intentional and actual post-inoculation talk, the next phase aims to examine the potential gap between intentions and actions in relation to post-inoculation talk. Lastly, with these initial results in mind, the next chapter aims to explore whether PIT is more than a vocalised form of counterarguing, and instead, a potential pathway to strengthening and spreading resistance.

# PHASE 2

# 5.4 Shedding light on PIT's role in attitudinal resistance

The purpose of this follow-up study was three-fold: to assess whether post-inoculation talk did occur organically, whether engaging in different forms of post-inoculation talk (sharing intentions, frequency and number of conversations, and degrees of resistance) has any effects on the individual, and whether there is a difference in intended post-inoculation talk and actual post-inoculation talk. That is, by asking participants to recall the content of their *actual* conversations, I distinguish between intended (Phase 1) and actual post-inoculation talk (Phase 2). To further optimise the scalability of inoculation interventions and to have a chance at outpacing the virality, pace, and depth at which online misinformation travels, this chapter also aims to examine whether the potential for 'super spreaders' of resistance exists. Indeed, Compton and colleagues (2020) have reflected on the advancements in the inoculation scholarship and have outlined new avenues for future research - at its core, the potential for establishing psychological 'herd immunity' against persuasion. Thus, this phase aimed to continue reassessing current conceptualisations of resistance and PIT and to establish the pre-requisites which must be met for post-inoculation talk to occur, be strong, and spread across networks. Consequently, for the purpose of clarity and brevity on the effects of spreading PIT on the spreader, the pre-registered hypotheses  $H_7$ ,  $H_8$ ,  $H_9$ , and  $H_{10}$  were combined (see below for updated hypotheses).

## Method

### Procedure

Participants from Phase 1 who were recruited through the crowd-sourcing platform *Prolific Academic*, were whitelisted and re-invited to take part in a follow-up study a week after initial treatment exposure. Therefore, participants completed the same post-test again, consisting of previously seen measures (counterarguing, motivational threat, apprehensive threat) and previously unseen measures of post-inoculation talk, ranging from frequency and depth (Ivanov et al., 2016) to more content-focused tasks asking to report *what* individuals talked about in the week between initial treatment exposure and the follow-up study. The latter was used to develop follow-up versions of the previously described measures of passing on inoculation, passing on misinformation, and a more general 'degrees of resistance index'.

### Measures

On top of the measures of sharing intentions and intentional pass-along messages reviewed in Phase 1, the following set of measures were introduced as new additions to measure PIT in Phase 2. Whereas the traditional PIT measure is taken from limited existing research (Ivanov et al., 2016), the remaining

measures were developed to provide a more nuanced picture of post-inoculation talk. For all PIT measures and scoring procedures, see Table 1.

#### Traditional PIT

Traditional PIT was derived from two self-reported measures following Ivanov and colleagues (2012): (1) by reporting the number of people each individual talked to about the fictious chemical in the interim between treatment exposure and the follow-up ("In one sentence, what did you tell others about the **common manipulation strategies**?") and (2) by including the frequency of conversations ("Since last week, **how many conversations** have you had about common *manipulation techniques* used to spread misinformation?"). In this study, these two questions concerning the quantity of talk were applied to both the spread of inoculation and misinformation. That is, participants were also asked to report how many conversations and how many conversation partners they have had about the **dangers of tryptostine**. Reliability analyses suggest good levels ( $\alpha$ = .84) of internal consistency (George & Mallery, 2003).

### Actual pass-along message

In a similar vein to the 'intentional pass-along message' measure ( $T_1$ ), participants were asked to write a short summary (up to 140 characters) on the content of the conversations they *actually* had in the interim between treatment exposure and the follow-up study. To achieve a more nuanced picture of whether misinformation or inoculation-specific aspects were passed on, participants were therefore asked "In one sentence, what did you tell others about the **common manipulation strategies**?" and "In one sentence, what did you tell others about *tryptostine*?" These responses were used to independently score for a series of PIT measures.

### Resistance through PIT

Scoring of pass-along messages depending on whether the messages spread basic resistance (1), inoculation (2), misinformation (-1), or neither/something irrelevant to the topic issue (0). The independent scoring of resistance in the pass-along messages at  $T_2$  is positively correlated (r = .91, p < .001).

#### 'Degrees of resistance'

Lastly, to provide a more nuanced account of attitudinal resistance, collected pass-along messages in phase 1 were used to develop a resistance index (see Phase 1, Table 1, for breakdown of index scoring; the supplements, for visualisation of the index). Accordingly, pass-along messages were scored on a scale from conspiratorial thinking (-3), spreading misinformation (-2), repeating misinformation (-1), sharing nothing (0), general resistance (1), enhanced resistance (2), and advocative resistance (3). The independent scoring of resistance in the pass-along messages at T<sub>2</sub> is positively correlated (r = .88, p < .001).

### Quality PIT

This measure was created by multiplying the number of conversation partners by participants' degrees of resistance score for their actual pass-along messages composed at  $T_2$ . Doing so aims to examine the 'super spreaders' of inoculation vs. misinformation. That is, high quality PIT will consist of actual pass-along messages that scored high on the 'degrees of resistance' scale (i.e., spread inoculation) and were shared with a high number of conversation partners whereas low quality PIT will consist of the lowest resistance score (i.e., spreading conspiracies) multiplied by the number of people they talked to.

# Results

It is worth noting that the present research operationalised post-inoculation talk on multiple dimensions (ranging from frequency and content to quality of talk). While doing so allows a more nuanced understanding than currently offered by inoculation research, it constraints most assessments of its effects on resistance (and vice versa) to more exploratory forms of analyses. Therefore, for the purpose of clarity and brevity, previously pre-registered hypotheses **H**<sub>7-10</sub> and additional exploratory analyses were conducted to examine the different facts of post-inoculation talk. Accordingly, the revisited hypotheses were the following:

 $H_7$ : Post-treatment talk will be positively associated with perceived resistance, attitudes, attitude strength, and attitude confidence towards the issue topic.

 $H_8$ : After a week, participants in the inoculation condition will score significantly higher on PIT scores (frequency and number of conversations, 'degrees of resistance', and quality of PIT) and spread less content from the attack message than participants in the control conditions.

# Role of PIT in the inoculation process

First, to assess whether PIT was positively associated with resistance<sup>35</sup>, attitudes, attitude strength, and attitude confidence towards the topic, a series of multiple linear regressions were run<sup>36</sup>. Considering the variety of PIT measures employed in this chapter (see Table X),  $\mathbf{H}_7$  was tested by focusing on sharing intentions (T<sub>1</sub>), the number of conversations (T<sub>2</sub>), and the quality of post-inoculation talk ('degrees of resistance' score x number of conversation partners; both T<sub>2</sub>). The results provide support for  $\mathbf{H}_7$ : Post-treatment talk predicts the degrees of resistance displayed in their actual pass-along messages, the attitudes towards the issue topic, attitude strength, and attitude certainty at the follow-up stage. They also point towards interactions between condition and quality of PIT which demonstrate that inoculated individuals spread higher quality PIT. See Table3 2 for summary of findings from the multiple linear regressions.

While the intentional pass-along message (T<sub>1</sub>) has offered initial insights into what people would share with others about online, a week after treatment exposure, participants completed a text box entry task asking them to summarise what they *actually* talked to others about in the interim between treatment exposure and follow-up ("In one sentence, what did you tell others about *tryptostine*? "). A One-way ANOVA suggests a significant main effect of condition on the spread of misinformation occurring in the talk in the interim between treatment exposure and the follow-up, F(1,251) = 13.87, p=0.001. Tukey's post-hoc comparisons show that individuals in the control condition shared significantly more content that spread misinformation about the issue topic than the inoculation condition ( $M_{control}= 0.13$ ,  $M_{inoc}=0.02$ ,  $M_{diff}=0.102$ ,  $p_{tukey}=0.001$ , 95% CI [0.04, 0.156] d= 0.39). Lastly, there is a significant main effect of condition on the quality of PIT apparent, F(1,318)=4.17, p=0.04,  $\eta^2=0.012$ . Tukey's post-hoc analysis shows that inoculated participants spread PIT that is significantly higher in quality (resistance score x number of conversation partners) than the control condition ( $M_{control}= 1.1$ ,  $M_{inoc}=1.75$ ,  $M_{diff}=-0.65$ ,  $p_{tukey}=0.042$ , 95% CI [-0.42, -0.007] d=-0.22). Additionally, a simple word cloud map (see Figure 29) demonstrates the most frequently used words occurring in actual pass-along messages. Though observational at this stage, there appear to be some interesting differences

 $<sup>^{35}</sup>$  Given the initial findings calling for a more nuanced approach to attitudinal resistance, this section will use the 'degrees of resistance' score of actual pass-along messages (T<sub>2</sub>) as an indicator of whether individuals demonstrated resistance against the attach message.

<sup>&</sup>lt;sup>36</sup> With these measures functioning as dependent variables in the multiple linear regressions.

compared to the map in Phase 1 (intentional pass-along message). For instance, whereas the most frequently used words in the first word cloud plot are "chemical", "dangerous", "information", and "research", the second plot demonstrates a shift towards words such as "manipulation", "manipulation strategies/techniques", "social media", "false information", "source of information", and "spread of misinformation". This pattern is consistent with the findings that post-treatment exposure, more inoculation-congruent content (forewarning, refutations) was passed on.

| DV  | Resistance |      |       |       | Attitudes |       |       | Attitude strength |          |      |       | Attitude certainty |          |      |       |       |
|---|------------|------|-------|-------|-----------|-------|-------|-------------------|----------|------|-------|--------------------|----------|------|-------|-------|
| Predictor                                     | Estimate   | SE   | t     | р     | Estimate  | SE    | t     | р                 | Estimate | SE   | t     | р                  | Estimate | SE   | t     | Р     |
| Intercent                                     | 0.5950     | 0.15 | 3.89  | <.001 | 3.73      | 0.15  | 25.57 | <.001             | 2.29     | 0.13 | 17.2  | <.001              | 2.69     | 0.23 | 11.57 | <.001 |
| Sharing intentions                            | -0.0621    | 0.03 | -1.97 | 0.049 | -0.07     | 0.03  | -2.23 | 0.026             | 0.18     | 0.03 | 6.62  | <.001              | 0.16     | 0.05 | 3.42  | <.001 |
| Number of conversations                       | 0.19       | 0.04 | 5.25  | <.001 | 0.02      | 0.04  | 0.53  | 0.596             | 0.15     | 0.03 | 4.65  | <.001              | 0.17     | 0.06 | 2.89  | 0.004 |
| Quality of PIT                                | 0.07       | 0.02 | 3.89  | <.001 | 0.07      | 0.017 | 4.1   | <.001             | -0.02    | 0.02 | -1.1  | 0.27               | -0.07    | 0.04 | -1.82 | 0.069 |
| control)                                      | 0.4        | 0.11 | 3.55  | <.001 | 0.08      | 0.11  | 0.76  | 0.445             | -0.06    | 0.1  | -0.64 | 0.52               | 0.31     | 0.17 | 1.82  | 0.070 |
| Quality PIT x Condition (inoculation-control) | -0.09      | 0.03 | -2.92 | 0.004 | -0.07     | 0.03  | -2.12 | 0.03              | -0.03    | 0.03 | 17.2  | 0.34               | 0.12     | 0.05 | 2.31  | 0.021 |
| Adjusted R <sup>2</sup>                       | 0.15       |      |       |       | 0.84      |       |       | 0.26              |          |      |       | 0.11               |          |      |       |       |

Table 2: Linear regressions measures of resistance, attitudes, attitude strength, and attitude certainty as the dependent variables (T2).

manipulation techniques on manipulation nothing nothing nothing argument lack of citacion anyone anything Subject false information opinion common manipulation strategy fake news conversatio researc modern day propaganda P different source favour of conspirancy )ne kein geheimnis und anvt much information many people everything everyday life manipulation social media spread of misinformation information last week chemical side effects common strategy manipulation strategy source of information misinformation group of people emotional manipulation

manipulative strategy kind of manipulation

### Figure 29: Word cloud map of frequently occurring words in actual pass-along messages (T2).

In sum, these results predominantly support  $H_8$ , while there was no difference in frequency of conversations or number of conversation partners between conditions, what inoculated individuals reported to have talked about qualified significantly higher on the 'degrees of resistance index' than talk by the control condition. Additionally, inoculated individuals spread significantly more high-quality PIT. That is, spreading content that scored highly on the 'degrees of resistance' index x number of people they conversed with. This offers crucial first insights into the potential role of inoculated individuals as 'super spreaders' of attitudinal resistance.

### Intention-action gap

Considering that in Phase 1 participants were asked to write down a message reflecting what they would share with others after treatment exposure and that Phase 2 asked them to report on what they in fact talked about in the interim, this study design allows to further examine the intention-action gap in post-inoculation talk. Accordingly, a series of exploratory analyses were run on to examine differences between conditions and any effects of time x condition on the spread of resistance vs misinformation through talk.

To begin with, a one-way ANOVA was conducted to assess whether PIT (specifically, the traditional account of the number of conversations and conversation partners) differed in the interim between treatment-exposure and follow-up between conditions (inoculation, control). The results show no

significant effect of condition on the frequency of talk, F(1,351) = 0.28, p = 0.59,  $\eta^2 = 0.001$  nor the amount of conversation partners, F(1,349) = 0.63, p = 0.42,  $\eta^2 = 0.002$ . Thus, these results are inconsistent with previous PIT literature in that inoculated individuals did not report to engage in more post-treatment talk than the control group.

# Spreading Misinformation

To assess whether the extent to which (intended vs. actual) pass-along messages spread misinformation changed over time, a repeated measures one-way ANOVA was conducted with condition (inoculation, control) as the between-subject factor and time ( $T_1$ ,  $T_2$ ) as the within-subject factor. Doing so illustrates a significant effect of time x condition on post-treatment messages spreading misinformation, F(1,349) = 8.93, p=0.003,  $\eta^2 = 0.009$ . Additional difference-in-difference analysis ( $M_{diffT2T1,control} = -0.2$ ,  $SD_{diffT2T1,control} = 0.5$ ;  $M_{diffT2T1,inoc} = -0.34$ ,  $SD_{diffT2T1,inoc} = 0.49$ ) using a post-hoc *t* test indicates a significant mean difference of  $M_{diff-diff} = 0.14$ , *t* (350) = 2.68, p = 0.008, 95% CI (0.03, 0.24), d = 0.28, showing that inoculated individuals spread significantly less misinformation about the issue through PIT (see Figure 30).



Figure 30: Line graph for spreading misinformation through pass-along messages between conditions (1= control; 2= inoculation).

# Spreading resistance

To test whether there is a main effect of condition (inoculation, control) on the difference of degrees of resistance displayed in the intended pass-along message  $(T_1)$  vs. the actual pass-along message  $(T_2)$ ,

a repeated measures ANOVA was run. Results suggest no significant main effect of condition x time on spreading resistance through PIT, F(1,349)=0.72, p=0.39,  $\eta^2=0.001$ . However, the results do suggest a significant effect of time, F(1, 349)=23.8, p=0.001,  $\eta^2=0.029$  and a significant effect of condition independently, F(1. 349)=15.8, p=0.001,  $\eta^2=0.023$ . Additional Tukey's post-hoc comparisons show actual post-inoculation talk (T<sub>2</sub>) by inoculated individuals spread significantly more resistance than intended talk (T<sub>1</sub>), ( $M_{T2,inoc}=1.02$ , SD<sub>T2,inoc</sub>=1.01;  $M_{T1,inoc}=0.15$ , SD<sub>T1,inoc</sub>=1.58 ;  $M_{diff}=0.87$ ,  $p_{tukey}=0.001$ , 95% CI [-0.12, -0.52]). Moreover, though no significant difference between conditions at T<sub>1</sub> is evident, when followed-up with a week later, inoculated individuals' report of their actual talk spread significantly more resistance than the control condition, ( $M_{T2,inoc}=1.02$ ,  $SD_{T2,inoc}=1.0$ ;  $M_{T2,control}=0.58$ ,  $SD_{T2,control}=1.17$ ;  $M_{diff}=-0.43$ ,  $p_{tukey}=0.009$ , 95% CI [-0.8, -0.07]). Thus, inoculated individuals spread significantly more resistance in their actual talk than in their intended post-inoculation talk while also spreading more resistance through PIT than the control condition (see Figure 31).



Figure 31: Line graph for spreading resistance through pass-along messages between conditions.

### Discussion – Phase 2

The follow-up study found evidence for  $H_7$ - $H_8$  and finds support for the role of PIT in conferring attitudinal resistance as well as predicting attitudes towards the issue topic, attitude strength, and attitude certainty. This suggests that PIT may not only be a form of vocalised counterarguing but rather, play an active role in strengthening the inoculation process. Additionally, though there was no difference in the number and frequency of conversations across conditions, the results point towards a noticeable difference between intended talk and actual post-treatment talk, further emphasising the need to establish a more nuanced picture of post-inoculation talk. In that vein, content analyses

illustrate that inoculated individuals' pass-along messages score higher on the 'degrees of resistance' index and that they share high-quality PIT. When assessing the degrees of resistance of PIT in relation to the number of conversation partners, inoculated individuals spread significantly more high-quality post-treatment messages. Thus, phase 2 of this chapter contributed to the ongoing scientific discussions about the effects of post-inoculation talk on the inoculated.

Research on the driving factors and consequences of PIT are inconsistent, resulting in two main accounts of why and how PIT occurs. Firstly, some research suggests that the induced state of threat and heightened awareness of one's susceptibility to persuasive attacks leads inoculated individuals to engage in assurance-seeking behaviours (Compton & Pfau, 2009; Pfau et al., 2003; 2004). That is, the 'incubation period' after treatment exposure makes the 'fragility' of one's attitudes more salient and leads individuals to talk about the inoculation treatment in order to make sense of, and thereby bolster, their attitudes (Parker et al., 2012b; Rosnow et al., 1988). On the other hand, some inoculation scholars believe that the heightened awareness of attitudinal susceptibility followed by counterarguing elicits advocating mechanisms (Compton & Pfau, 2009; Petty & Wegener, 1998; Pfau et al., 2000). Accordingly, this notion suggests that inoculated individuals, sometimes reinforced by anger about the issue, feel motivated to share what they have learned with others. Thus, underpinning PIT with factors of attitudinal confidence, a greater sense of responsibility, and advocacy (Compton & Pfau, 2009). Though these two approaches offer diametrically opposed explanations as to why inoculated individuals do participate in talk post-treatment exposure, Ivanov and colleagues (2015) argue PIT to be multifunctional and instead, is capable of serving as reassurance and advocacy to bolster resistance. However, it is worth noting that the few existing empirical studies on PIT predominantly employed instructions to engage in talk post-treatment exposure. For instance, Ivanov and colleagues (2015) followed up with participants and instructed them to share a specific and detailed written account of their most in-depth conversation about the issue topic. Similarly, Ivanov and colleagues (2018) asked half of their sample to not share or discuss anting in relation to the issue topic post treatment-exposure, while the other half was instructed to do the opposite.

Since the present study did not examine the motivations behind PIT, the results do not allow to make conclusions about the functions and motivations of PIT in the inoculation process. Instead, this follow-up study aimed to assess whether PIT occurred organically, whether it functioned as a type of booster treatment on the spreaders' attitudinal resistance, and whether the talk spread more inoculation-congruent content than misinformation. Although the present findings relied on selfreported recalling of post-inoculation talk, instead of talk *itself*, capturing that actual PIT spread significantly more resistance and inoculation-congruent content than intended talk, shed unique insights into the intention-action gap of inoculation-driven behaviours. These results offer insights into the dimensions of post-inoculation talk, ranging from intentions to talk about inoculation content immediately after treatment exposure, the actual number of conversations and conversation partners, and what these individuals in fact talked about.
By establishing a lexicon for post-inoculation talk, the messages were mapped onto a spectrum which dismantles the dichotomised notion of fake and true, and instead, calls to re-think resistance by paying closer attention to the 'degrees of resistance'. This index ranges from contributing to and spreading conspiracy theories about the issue topic to actively spreading forewarnings and second-order refutations about the manipulation techniques underpinning the misinformation message. Doing so sheds more light on the inoculation process and how resistance is built, strengthened, and spread through talk. Accordingly, the findings show that inoculated individuals' pass-along messages are more advocative and congruent with spreading inoculation even when pitted against misinformation. In other words, although no *quantitative* differences in talk between the inoculation and control condition are apparent, the content-scoring of the pass-along messages exemplifies that inoculated individuals spread more resistance (i.e., pre-emtoive refutations about manipulation techniques) and less misinformation about the issue topic than the control group. The significant effect of condition on the quality of PIT also demonstrates that inoculated individuals who scored higher on the resistance index spoke to more individuals in their (social) network about the issue topic. Nevertheless, future research should investigate ways of capturing PIT as it unfolds and spreads online.

Importantly, this novel design brings forward several theoretical explorations. First, the attack message focused on the issue topic (the fictitious chemical, tryptostine) whereas the inoculation message offered a generalised account of manipulation techniques underpinning common health misinformation content. While the three strategies mentioned in the treatment message (fearmongering, use of fake experts, and conspiracy theories) matched the ones employed in the attack message, they did not mention the fictitious chemical (or any other chemicals) in their refutational pre-emptions (see Appendix for all treatment messages). In doing so, this phase adds to two contemporary theoretical foci that remain largely underexplored.

As initially theorised (Papageorgis & McGuire, 1961), questions around the generalisability and required specificity of inoculation treatments has resulted in a set of different mechanisms that allow for the extension of attitudinal resistance against unmentioned arguments on the same topic (blanket of protection) as well as related yet untreated topics ('cross-protection'). On top of that, the scarce literature mentions 'umbrella of protection', which appears to be used interchangeably with 'blanket of protection' (Parker et al., 2012b, 2016). This calls for inoculation scholarship to prioritise the terminological and theoretical clarity around spreading intra- and inter-individual resistance. In other words, from studying the spill-over of resistance to unmentioned arguments within the same topic and conferring resistance to untreated yet related topics on the one hand, to spreading resistance from one individual to another, current scientific conceptualisation does not have a clear understanding of the boundary conditions or prerequisites needed to extend resistance to different arguments, topics, and individuals. In this study, the inoculation treatment incorporates foundational similarities (manipulation techniques) on different, untreated, and varyingly related topics (all on health yet not necessarily attitudinally associated) and thereby, offer an 'umbrella of protection' that is both specific and generalised. One that seems to incorporate the efforts of a 'blanket of protection' by providing refutational pre-emption against manipulation techniques while also conferring 'cross-protection' against a previously unseen and untreated topic issue. With that in mind, one could propose 'umbrella of protection' to be used for inoculation treatments that are generalised enough to be applied against unrelated topics using the same manipulation techniques.

These are novel insights into the theoretical and applicable boundaries of inoculation, contributing to its potential extension from the conceptually separate research efforts of (1) testing issue-specific inoculation treatments on attitude-congruent topics and (2) assessing generalised inoculation treatments that focus on underlying manipulation mechanisms within the same contextual sphere towards a more hybrid form. One which is simultaneously conferring resistance against common manipulation techniques while applying them to untreated and unrelated topics (e.g., side-effects of earbuds). Early research provides support for the notion that changing one belief can affect related beliefs (McGuire, 1964). Indeed, research on the Galileo spatial-linkage model of inter-attitude structure and dynamics (Dinauer & Fink, 2005) as well as the hierarchy of internal processes and attitudes ((Hunter et al., 1976) suggests that when related attitudes change, the cognitive dissonance resulting from it can trigger a ripple effect where associated attitudes are adjusted to the changes in the targeted attitudes (e.g., vaccines) – suggesting that such 'umbrella of protection' efforts could be tailored towards less crystallised yet associated attitudes on the periphery to extend attitudinal resistance against a number of attitudes.

Arguably, theoretical room needs to be made for these contributions to be accommodated into a theory for which its scholarship has predominantly constrained itself within the possibilities of the biological analogy it is based on. However, Compton (2013) suggests that the analogical foundations of inoculation theory should be regarded as instructive rather than prescriptive, a notion that would allow a broader conceptualisation of inoculation-based resistance. Similarly, instead of falsely separating the internal and external communication part of counterarguing, another theoretical tenant, more research is needed to assess how PIT fits the analogy. By incorporating several operationalisations of PIT, phase 2 makes substantial contributions to the current scientific understanding of whether, when, and how much post-inoculation talk occurs organically and how, in turn, it effects the spreader. Specifically, the results suggest that PIT predicts perceived ability to spot and refute misinformation about the issue topic as well as the certainty and strength with which these attitudes are held, highlighting its strengthening role in the inoculation process. Moreover, insights into the content of PIT suggests that PIT may be viewed as more than a vocalised form of counterarguing, and instead, may play a substantial role in the spreading of resistance, too. While these are noteworthy contributions, they do not allow to draw any conclusions about whether PIT has any effects on the spreader nor whether passing on second-order inoculation messages confers resistance to the 'vicariously inoculated', the recipients of PIT. Thus, arguably the most promising aspects and potential of PIT, that is, its use for conferring inter-individual attitudinal resistance remains unexplored.

# PHASE 3

# 5.5 Interindividual resistance through vicarious inoculation

Thus far, this Chapter has offered multiple insights into post-inoculation talk. To summarise, Phase 1 and Phase 2 explored whether inoculated individuals, voluntarily and without any instructions, engage in talk after treatment exposure. Specifically, while no differences in quantity and frequency of conversations across conditions is evident, inoculated individuals spread more inoculation-congruent content and less misinformation about the chemical in their actual talk. Furthermore, the findings suggest that PIT plays an important role in the strengthening and spread of inoculation by predicting perceived resistance against misinformation, attitude strength, and attitude certainty. In this next part, the aim is to shift the focus from intraindividual resistance to interindividual resistance. To do so, this phase will examine whether, when pitted against misinformation about the issue topic, second-order PIT, that is, actual pass-along messages composed by different participants in Phase 2, prevail. In other words, is it possible to vicariously inoculate individuals through PIT? And lastly, when exposed to both PIT and misinformation, what do these recipients choose to pass on to others? To summarise, this phase will examine whether PIT can function as a substitute for inoculation treatments and confer vicarious resistance to the recipients of talk. Though this far this chapter has approached PIT as an organic occurrence and extension of vocalised counterarguing (Phase 1) and a process of inoculation (Phase 2), this last phase pivots towards thinking about PIT as an outcome of inoculation and a potential pathway towards psychological 'herd-immunity' against online misinformation.

# Methods

### Sample

A total of 600 participants (200 per condition) were recruited through the crowd-sourcing platform, *Prolific Academic*. Given that, at the time of this write-up, no study has explored the efficacy of postinoculation messages as sole treatment stimuli, the final sample size was estimated. Additionally, all participants from Phases 1 and 2 were blacklisted and beyond being fluent in English and being above 18 years old, no other exclusion criteria were implemented. Of the sample, 50.1% identified as female (48.2% male, 1.6% other); 86.3% indicated being between 18 and 34 years of age, and 43% reported having a bachelor's degree. Lastly, the sample skewed politically left (M = 3.03, SD = 1.34).

# Procedure

Participants were recruited through the online crowd-sourcing platform, Prolific Academic and were first asked to indicate their attitudes towards the issue topic (fictitious chemical, tryptostine), their attitude certainty, and their perceived resistance against arguments about the dangers of the chemical. Subsequently, participants were randomly assigned to one of three conditions (inoculation, misinformation, control). Though the design followed the previous two studies, instead of the treatment messages, participants were exposed to a set of tweet-sized messages and were told that these were shared by a previous round of participants with them. These consisted of the post-treatment messages composed by participants in Phase 1. More specifically, of these messages, 7 which scored highest/lowest on the 'degrees of resistance' index were used to compose a newsfeed-stimulating set of inoculation/misinformation PIT tweets, respectively. Therefore, the inoculation condition was exposed to 7 inoculation messages followed by 7 misinformation messages, while the misinformation condition only received the latter, and the neutral condition was shown 7 made-up and unrelated messages (see Appendix). Afterwards, all participants were required to complete two threat measures (apprehensive: motivational), indicate their attitudes, attitude certainty, sharing intentions, perceived ability to resist and counterargue manipulative content about the issue topic. Lastly, all participants were asked to compose a message for the respective next round of participants ("In one sentence, what would you tell the next participants about the chemical tryptostine?"). Upon completion, all participants were debriefed and received an additional e-mail about the fictitious nature a few days later. This study was approved by the Cambridge Ethics Committee (PRE.2022.006).

## Materials

The treatment messages were taken from the messages of the intended PIT composed in Phase 1. More specifically, after treatment exposure, all participants were asked to leave a tweet-sized message about what they would like to share with others. These messages were independently reviewed by two researchers and were given a 'degrees of resistance' score. The two ends of this spectrum meant a score of -3 for PIT messages that actively contribute to and spread conspiratorial thinking about the issue topic to +3 for content that is inoculation-congruent and incorporated elements of passing on threat, forewarning, and pre-emptive refutations. Accordingly, 7 of the highest and lowest scoring messages from Phase 1 were used to develop PIT-based and newsfeed-stimulating treatment messages for Phase 3 (see Figure 32 for examples)



Figure 32: Examples of actual pass-along messages (T2) turned into Tweet-sized treatment material (left to right; inoculation, misinformation, control condition).

## Measures

All measures used in phase 3 have previously implemented in phases 1 and/or 2. These include measures of perceived resistance, attitudes, attitude certainty, counterarguing, and threat (apprehensive, motivational), the same PIT measures from phase 2 were used (sharing intentions, text box entry). Additionally, identical to phase 2, messages composed by participants in phase 3 were scored by two independent researchers and resulted in content-based measures of post-inoculation talk (sharing intentions, intentional PIT). However, two separate measures allowed to score the content of intentional PIT.

#### Spreading misinformation

Building on the codebook developed for phase 2, the composed messages were independently scored based on whether they passed on misinformation about the issue topic (0= no sharing of misinformation content; 1= repeating or adding onto misinformation content). The independent scoring of resistance in the pass-along messages at T<sub>2</sub> is positively correlated (r = .69, p < .001).

#### Spreading inoculation

Similarly, the established codebook was used here to independently score whether composed messages pass on inoculation content (0=no mention; 1= basic resistance; 2= inoculation congruent). The independent scoring of resistance in the pass-along messages at T<sub>2</sub> is positively correlated (r = .80, p < .001).

### Results

Considering that, at the time of the write-up, there is no existing research on the efficacy of secondorder inoculation messages, the analyses in phase 3 were exploratory. Thus, this phase aimed to explore whether 1.) vicarious inoculation can confer resistance against the issue topic, 2.) whether inoculated individuals spread less misinformation about the issue topic than the other conditions, and 3.) whether receiving an inoculation treatment vicariously has any impact on the attitudinal ascendents (e.g., attitude certainty, strength, counterarguing).

# Attitude change

To assess whether individuals' attitude change differed between experimental conditions, a one-way ANOVA analysis was conducted. The results demonstrate a significant main effect of condition on attitude change, F(2, 395) = 4.44, p=0.01,  $\eta^2 = 0.015$ . Tukey's post-hoc analyses show that inoculated individuals experience significantly less attitude change than the control condition ( $M_{control} = -0.5$ ,  $M_{inoc} = -0.22$ ,  $M_{diff} = -0.28$ ,  $p_{tukey} = 0.008$ , 95% CI [-0.49, -0.1] d = .29). Interestingly, there is no significant difference in attitude change between inoculated individuals and the misinformation condition ( $M_{misinfo} = -0.39$ ,  $M_{inoc} = -0.22$ ,  $M_{diff} = -0.168$ ,  $p_{tukey} = 0.17$ , 95% CI [-0.37, -0.01] d = .18).



Figure 33: Violin plot with jitter of attitude change (post-pre) between conditions.

# Perceived Resistance

Next, a one-way ANOVA demonstrates a significant main effect of experimental condition on perceived resistance, F(2, 594)= 5.99, p= 0.004,  $\eta^2 = 0.019$ . More specifically, Tukey's post-hoc analyses show that inoculated individuals experience significantly less difficulty coming up with arguments against dangers of the fictitious chemical than the control condition ( $M_{control}= 4.46$ ,  $M_{inoc}=3.95$ ,  $M_{diff}=0.52$ ,  $p_{tukey}=0.002$ , 95% CI [0.13, 0.53] d= .337). On the contrary, there was no significant difference between the inoculation condition and the misinformation only condition apparent ( $M_{misinfo}=4.19$ ,  $M_{inoc}=3.95$ ,  $M_{diff}=0.24$ ,  $p_{tukey}=0.25$ , 95% CI [-0.03, 0.35] d= .16). See Figure 34 for visualisation of perceived resistance (T<sub>2</sub>) across conditions.



Figure 34: Violin plot with jitter of perceived resistance scores (T2) between conditions.

# Counterarguing

Similarly, a one-way ANOVA shows a significant main effect of condition on individuals' selfreported ability to counterargue against misinformation messages about the chemical, F(2, 594)=4.92, p=0.008,  $\eta^2 = 0.016$ . Additional Tukey's post-hoc analyses are consistent with findings above in that inoculated individuals experienced significantly greater ability to refute misinformation messages about the fictitious chemical than the control condition ( $M_{control}=3.87$ ,  $M_{inoc}=4.33$ ,  $M_{diff}=-0.46$ ,  $p_{tukey}$ = 0.006, 95% CI [-0.5, -0.11] d=.31) while no significant difference is evident between the inoculation and misinformation only condition, ( $M_{misinfo}=4.15$ ,  $M_{inoc}=4.33$ ,  $M_{diff}=-0.18$ ,  $p_{tukey}=0.45$ , 95% CI [-0.31, 0.07] d=.12). See Figure 35.



Figure 35: Violin plot with jitter of self-reported counterarguing (T2) between conditions.

# Spreading misinformation

Finally, a one-way ANOVA was run to test whether there was a difference between conditions on whether participants' pass-along messages spread misinformation about the fictitious chemical or not. Results demonstrate a significant main effect of condition on spreading of misinformation, F (2, 594) = 12, p= .001,  $\eta^2$ =0.039. More specifically, Tukey's post-hoc analyses shows that inoculated individuals spread significantly less misinformation than the control ( $M_{control}$ = 0.43,  $M_{inoc}$ =0.22,  $M_{diff}$  = 0.2,  $p_{tukey}$  = 0.001, 95% CI [0.27, 0.67] d= .47) and the misinformation only condition ( $M_{misinfo}$ = 0.37,  $M_{inoc}$ =0.22,  $M_{diff}$  = 0.15,  $p_{tukey}$  = 0.002, 95% CI [0.15, 0.54] d= .35). Lastly, there was no significant difference between the control and misinformation condition in terms of the pass-along messages qualifying as spreading misinformation ( $M_{control}$ =0.43,  $M_{misinfo}$ = 0.37,  $M_{diff}$  = 0.06  $p_{tukey}$  = 0.41, 95% CI [-0.07, 0.33] d= .13).



Figure 36: Violin plot with jitter of post-treatment messages that spread misinformation between conditions.

# Spreading inoculation

Lastly, a one-way ANOVA illustrates a significant main effect of condition on whether participants' pass-along messages spread inoculation, F(2, 594) = 28.1, p = 0.001,  $\eta^2 = 0.086$ . Tukey's post-hoc analyses shows that inoculated individuals spread significantly more inoculation than the control  $(M_{\text{control}} = 0.53, M_{\text{inoc}} = 0.96, M_{\text{diff}} = 0.2, p_{\text{tukey}} = 0.001, 95\%$  CI [-0.89, -0.49] d = .69) and the misinformation only condition ( $M_{\text{misinfo}} = 0.59, M_{\text{inoc}} = 0.96 M_{\text{diff}} = -0.37, p_{\text{tukey}} = 0.001, 95\%$  CI [-0.79, -0.39] d = .59). Lastly, there was no significant difference between the control and misinformation condition in terms of the pass-along messages qualifying as spreading misinformation ( $M_{\text{control}} = 0.53, M_{\text{diff}} = -0.06 p_{\text{tukey}} = 0.57, 95\%$  CI [-0.29, 0.09] d = .10).



Figure 37: Violin plot with jitter of post-treatment messages that spread inoculation between conditions.

# Discussion – Phase 3

For over five decades, the effects of inoculation treatments have been conceptualised as an individual and internal process to the direct recipient of the inoculation message. More recently, research has proposed the notion of socially diffused inoculation through interpersonal conversations (Compton & Pfau, 2009; Dillingham & Ivanov, 2016; Ivanov et al., 2015). However, these studies have focused on the benefits of post-inoculation talk on the recipient, neglecting the potentially much more extensive effectiveness and applications of collective inoculation-based resistance. In contrast, the overarching aim of this doctoral thesis is to establish how inoculation is build, strengthened, and spread societally. Accordingly, Phase 3 sheds light on the potential to spread attitudinal resistance to misinformation on an inter-individual level. By exploring the efficacy of PIT messages in replacement of traditional inoculation treatments, phase 3 provides initial and promising results for vicarious inoculation. Additionally, the findings suggest that the vicariously inoculated individuals, too, spread less misinformation and more inoculation-congruent content in their respective tweet-sized pass-along messages. This chapter also sheds further light on the "umbrella of protection" conferred through inoculation, given that individuals were initially inoculated against manipulation techniques applied to untreated and unrelated topics within the overarching theme of health, yet displayed resistance against an attack message on a previously unseen topic. This generalised resistance worked for the vicariously inoculated individuals, emphasising the need to further explore the assumed theoretical and practical boundaries of inoculation treatments. At the time of the write-up, the first to explore the efficacy of vicariously received inoculation. Though the research was predominantly of exploratory nature, substantial initial findings have emerged from it, shedding light on the overlooked and potential role of post-inoculation talk in establishing psychological herd immunity through inter-individual attitudinal resistance to persuasion.

In this study, participants in the inoculation condition were exposed to a series of tweet-sized messages from previously inoculated participants. The findings provide initial support for conferring resistance through vicarious inoculation. By demonstrating that inoculated individuals experience significantly less attitude change, higher perceptions of resistance, and ability to refute misinformation about the issue topic than the control condition, initial findings suggest that post-inoculation talk can serve as an effective substitute for inoculation treatments. However, the results also highlight no significant differences between inoculated individuals and those in the misinformation only condition for these three traditional conceptualisations of resistance. Which is why the additional insights form the content analyses provide an arguably previously uncaptured portion of the inoculation process, one that challenges the current portrait of resistance. Interestingly, the PIT messages from Phase 1 that scored highest on the 'degrees of resistance' index all implemented the techniques covered in the treatment message to the novel, untreated, and unrelated topic of the fictitious chemical. In other words, participants resist, warn, and refute in their PIT messages by contextualising and tailoring it towards a different topic. Indeed, this is consistent with research proposing that PIT is not restricted to the treatment material and, instead, can extend beyond arguments raised in the treatment message (Banas & Rain, 2010; McGuire, 1964; Ivanov et al., 2015). While these findings are mostly observational, it is supported by inoculated individuals' messages scoring significantly higher on the resistance index and spreading more inoculation-congruent content than other conditions. However, one limitation of this study is in light of earlier theorising that having one's PIT challenged can result in weakened resistance (Compton & Pfau, 2019; Ivanov et al., 2012). Consequently, the present study does not allow to observe the reciprocal dynamic of post-inoculation talk and the effects of questioning and feedback on the inoculation process for the spreader and the recipient.

Instead, the present study took seven recalled post-inoculation messages that scored highest on the 'degrees of resistance' index and presented them to a new set of participants in place of a traditional inoculation message. By doing so, it can be argued that it does not adequately imitate an in-person conversation but rather anonymously presents elements of *recalled* talk to a new set of recipients. A large body of research on agent characteristics, persuasive messengers, and source credibility gives reason to assume that an in-person conversation spreading PIT would look differently (Touré-Tillery & McGill, 2015; Traberg & van der Linden, 2022), emphasising the need for future research to capture PIT in real-time, including its effect on the 'vicariously inoculated'. However, given that this is, at the time of the write-up, the first empirical study to examine whether interindividual inoculation through post-inoculation talk is possible, the present findings offer important foundational building blocks for future research to come. To begin with, it can be argued that this setup imitates an open social network (e.g., Twitter) where an individual can come across information that is limited by character counts and not necessarily directly composed or spread by their own social network. Thus, this trade-off allowed the present research to examine whether, when pitted against online misinformation, post-inoculation talk has the potential to keep up with, if not outpace online misinformation.

Though vicariously inoculated individuals did not differ in their perceived resistance, ability to refute manipulative arguments, nor in their attitudes towards the issue topic compared to the other conditions, the present study shows that when asked as to what they would like to pass onto the next round of participants, inoculated individuals spread significantly less misinformation about the issue topic and significantly more content that spread resistance than the control condition. Moreover, the findings suggest that inoculated individuals experience significantly less attitude change, higher perceived resistance, and increased ability to refute misinformation about the issue topic than the control condition. Despite the absence of any attitude change, traditionally used as a litmus test for inoculation-based attitudinal resistance, honing in on the content of third-order pass-along messages demonstrates that vicariously inoculated individuals also spread significantly less misinformation and more inoculation-congruent content about the issue topic. Yet, there was no difference in attitude certainty nor motivational and apprehensive threat amongst all conditions. This is contrary to findings in previous chapters where both attitude certainty and motivational threat are predictive of attitudinal resistance. One possible explanation is that second-order inoculation treatments do not elicit the same level of threat regarded as fundamental to the inoculation process, one could argue that it might be due to the format of the inoculation treatment. Though some research exists on the treatment modality of inoculation treatments, such as print vs. videos, (Lim & Ki, 2007; Pfau et al., 2000), this is the first to reduce a traditionally elaborate and extensive account of refutations to a series of anonymised Tweetsized messages composed by previously inoculated individuals, warning others about an issue that recipients could not have had any pre-existing attitudes towards. However, even though the traditional inoculation conditions may not be in place, research on the effectiveness of inoculation treatments on individuals with differing pre-existing beliefs (i.e., supportive, neutral, and opposed) gives reason to surmise that these PIT messages may still serve as effective treatments (Wood, 2007; Banas & Rains, 2010). On the other hand, while gamified inoculation treatments have provided support for active treatments that require the participants to come up with their own counterarguments, even Go Viral!, the shortest version to date (approximately 5 minutes) is still substantially longer than the treatment used here. Thus, Tweet-sized inoculation messages may not be able to compete with and result in mechanistic shifts as observed w traditional inoculation treatments do. As a result, phase 3 is unable to extend mechanistic insights established in phase 2 to the vicariously inoculated.

However, the current set up allows to assess the efficacy of inoculation treatments in a setting that is realistic to how individuals seek, come across, and make sense of (health) information today (Soroya et al., 2021; Zhao & Zhang, 2017). That is, on their newsfeeds – a space where news is presented in conjunction with seemingly harmless content of their social networks and yet, is underpinned by algorithms, filter bubbles, and echo chambers (Flaxman et al., 2016; Geschke et al.,

2019; Rhodes, 2022). Thus, by pitting character-limited inoculation messages against a set of equally short messages spreading misinformation about the issue topic, the current study examines the efficacy of such treatments in a more ecologically valid environment. Additionally, one might argue that in the contemporary state of information overload, it is not necessary for individuals to have fully formed opinions and attitudes towards each topic. Rather, the focus should be that when individuals come across potentially harmful content that employs manipulation techniques, they possess the ability to spot and resist spreading it further. Therefore, with the aim to examine the efficacy of (online) postinoculation talk when pitted against misinformation, the findings in phase 3 are more than promising. Consequently, these results highlight the need to rethink resistance and the way it is operationalised in inoculation research today. Moreover, much like research on the diffusion chain paradigm investigating how online information changes from one recipient to another recipient emphasised the role of pre-existing attitudes when receiving and passing on information (Giese et al., 2020), future research should investigate the distortion as well as the ripple-effect of post-inoculation talk. Advancing the scientific understanding of the role of PIT in the inoculation process and understanding when vicarious inoculations stop being effective constitutes as a crucial step towards enhancing and utilising inoculation-based resistance to its full potential.

#### 5.6 General discussion

The three studies in the present chapter aimed to address the theoretical gaps in the limited PIT literature by establishing, (1) whether post-inoculation talk occurs organically, (2) whether engaging in it has any effects on the spreader, and (3) whether receiving PIT can vicariously confer resistance to its recipients. In short, the findings provide support for all three statements. Specifically, the findings suggest that inoculated individuals do voluntarily engage in post-inoculation talk without any prior instructions, that PIT plays a role in conferring resistance, strengthening attitudes, and bolstering attitude certainty, and lastly, that PIT messages can effectively confer vicarious resistance to its recipients. In establishing these effects, the current chapter carefully portrays an updated conceptualisation of PIT. More specifically, by treating it first as a form of vocalised counterarguing, then a process of inoculation, and lastly, as a product of inoculation, this chapter points towards a path in which PIT can be effectively leveraged toward psychological herd-immunity against online misinformation. These are promising results and call for inoculation research to further investigate the boundaries and potential of post-inoculation talk, especially in the ongoing fight against harmful online misinformation.

# Revisiting and updating counterarguing

Until recently, inoculation scholarship considered one of its theoretical pillars, counterarguing, as a distinctly subvocal process. As a consequence, research focused on internal processes and individual differences that mediate such intrapersonal communication (Compton & Pfau, 2009). Yet, McGuire

(1964) included the notion of increased post-inoculation talk (PIT) in the early stages of his theorising. The present chapter suggests that post-inoculation talk predicts attitudes towards the issue topic, attitude strength, attitude certainty, and perceived resistance a week after treatment exposure. This is consistent with recent research, where Compton and Pfau (2009) argued that the unsettling nature of threat elicited by inoculation treatments might motivate individuals to talk to others about the issue, and inadvertently, strengthen their resistance and "function as a type of booster treatment" (Ivanov et al., 2012, p.26). The authors theorised that inoculation treatments could spread through interpersonal conversations, a proposition that was first empirically tested by Ivanov and colleagues (Dillingham & Ivanov, 2016; Ivanov et al., 2012; 2015; 2018). This initial research provided support for the notion that counterarguing is not only subvocal, but that inoculation-elicited processes lead to interpersonal word-of-mouth conversations. Furthermore, Ivanov and colleagues (2012) found that the frequency and amount of post-inoculation talk were positively correlated with resistance, that is, greater levels of talk lead to stronger beliefs about the target issue. Indeed, research suggests that inoculated individuals are more likely to speak about contested issues (Lin & Pfau, 2007), that the intrapersonal talk can be as equally effective as the treatment message itself (Southwell & Yzer, 2009), and that PIT has a positive impact on sharing issue-relevant information while also boosting attitude certainty (Dillingham & Ivanov, 2016). In sum, the combination of motivated defence, refutational pre-emption, and counterarguing practices is believed to equip the individual with the skills necessary to defend their attitudes against future persuasive attacks (Compton, 2013; Ivanov, 2012; 2017; McGuire, 1964).

It is important to note, that to adequately address these questions, the present chapter operationalised PIT in a critically different way to existing research. More specifically, so far, PIT was measured by giving participants instructions on whether or not to engage in talk and then by asking inoculated participants to report the number and amount of conversations they have had in the interim. On top of these measures, this series of studies employ a multi-faceted exploratory account of post-inoculation talk by including any instructions, asking about participants' intentions to talk about the issue, and requiring them to compose their PIT messages. These messages were character-limited, imitated posts made on social media platforms (i.e., Tweet-sized messages on Twitter), and measured intended talk immediately after treatment exposure as well as actual talk that organically occurred in the interim between phases 1 and 2.

These insights into the content of (intended and actual) talk allowed a more nuanced picture of what PIT is about. Namely, by coding the content of the messages and situating it on the 'degrees of resistance' spectrum, the present findings call for inoculation scholarship to rethink attitude resistance and post-inoculation talk in substantially novel ways. Though this chapter has provided further support for the notion that PIT has a strengthening effect on the *spreader*, this research takes an initial and novel step further by attempting to paint a more nuanced picture of post-inoculation talk, its effect on the mechanisms underpinning inoculation, and importantly, its effect on the recipient. Hence, by pairing the 'degrees of resistance' with the traditional measures of talk quantity, the findings in this chapter provide insights into the spread of high-quality post-inoculation talk. In other words, it takes an initial step towards using inoculated individuals as 'super spreaders' of PIT that can vicariously inoculate others, and thereby, contribute toward psychological herd immunity.

# It is all in the details – toward a more nuanced portrait of PIT

The findings of this chapter are largely not consistent with previous, albeit limited, research on postinoculation talk. However, neither are they entirely opposed but rather, call for the inoculation scholarship to rethink some of its core components. To begin with, these results are not consistent with previous research findings on inoculation treatments leading to *quantitatively* more talk. One reason might be that previous research employed instructions about whether or not participants should be talking to others in the interim between the treatment and attack exposure (Ivanov et al, 2012; 2015). In contrast, in this chapter, participants received no such instructions, allowing us to investigate whether inoculated participants do talk more *organically*. It can be argued that an exclusive focus on the quantity of talk is not adequately aligned with how information is retrieved, spread, and shared today (Gil de Zúñiga et al., 2017; Park, 2019). Within the context of inoculation theory, it is questionable whether the quantity of post-inoculation talk is the most scientifically adequate and relevant way of capturing and applying inoculation-induced talk.

It could be argued, however, that the absence of the optimal delay between treatment and attack message resulted in conditions not differing in their frequencies of conversations. In fact, establishing the optimal delay between treatment exposure and the subsequent attack has drawn much interest from inoculation scholars (e.g., Compton & Pfau, 2005; Ivanov, Pfau, & Parker, 2009), yet it was not a focal point of interest in this chapter. Though no quantitative difference in the talk was apparent, inoculated individuals did perceive higher levels of motivational threat, which, in turn, predicted perceived resistance and sharing intentions to talk about the treatment message with others. Considering the unpredictable exposure to online misinformation, this research opted for a short delay. While the meta-analysis finds no effect of delay on inoculation talk messages against misinformation occurring at different delay periods.

# Extending the reach of resistance

Indeed, by demonstrating the efficacy of inoculation treatments that do not cover the same issue topic or arguments as the attack message, this chapter calls for inoculation scholarship to rethink the boundary conditions of resistance through inoculation. Indeed, though the dominant use of attitude resistance is as an outcome, it is important to note that the absence of attitude change can be obtained without a resistance motivation driving it (Tormala et al., 2004). Instead, zero attitude change may be due to one's failure to attend to or comprehend the treatment or attack message (Hovland, Janis, & Kelley, 1953). Instead, some early research has proposed to think of resistance as a quality, where some types of people or attitudes are more resistant to change than others (Briñol et al., 2004; Petty &

Krosnick, 1995). Especially with the aim to develop inoculation treatments that could keep up with or even outpace online misinformation, restricting PIT to quantities limits its scientific understanding and application. Put differently, there is no need for unnecessarily constraining the potential of spreading resistance through talk, and instead, exploring its role in conferring an 'umbrella of protection' (Ivanov et al., 2012). Indeed, I propose that amidst the often interchangeably used terminologies for extensive attitudinal resistance (e.g., 'blanket of protection', 'cross-protection'), an 'umbrella of protection' may be an appropriate way to distinguish between 'blanket of protection' (i.e., resistance against different arguments within the same topic) and 'cross-protection' (i.e., resistance against untreated yet related topics). Instead, this chapter demonstrates the efficacy of an inoculation treatment that is both generalised and applied to an untreated and unrelated topic – suggesting that an 'umbrella of protection' against misinformation is possible. In other words, it appears that post-inoculation talk as a substitute of inoculation treatment messages is sufficiently specific to confer generalised immunisation against manipulation techniques, even if these are applied to the context of an untreated and unrelated topic. Future research should employ diffusion paradigms to further explore the ripple effect of such vicarious inoculation. Or put differently, similarly to the gradual distortion of messages in 'Telephone game', how many iterations of post-inoculation talk can successfully confer resistance from one individual to another before its content no longer meets the pre-requisites?

Situating the *actual* post-inoculation talk on the 'degrees of resistance' spectrum allowed to disentangle the false dichotomy of true vs. false content and instead, offers a more ambiguous assessment of when inoculation vs. misinformation is spread. In the same way as a sole focus on the frequency of PIT is not sufficient, having participants pass on messages that score high on the 'degrees of resistance' scale is not enough to establish psychological herd immunity if they only talk to one person. Hence, examining the quality of post-inoculation talk also showed that inoculated individuals spread significantly more high-quality messages with more people and significantly less misinformation with fewer people, respectively. Lastly, by using a series of messages from either end of the spectrum (spreading conspiratorial thinking; spreading threats and pre-emptive refutations), this study adapted the traditional modality of treatment and attack messages to the context of misinformation on social media (i.e., tweet-sized PIT). Importantly, considering that the treatment messages pre-emptively debunked (prebunked) manipulation techniques used on untreated and unrelated topics than the attack message, these findings also suggest that receiving treatments vicariously confers generalised resistance against manipulation. Overall, this constitutes the first study to examine and demonstrate the possibility of leveraging organic post-inoculation talk to confer vicarious inoculation. In sum, this chapter connects previously established driving factors such as attitude certainty and motivational threat and extends them to the context of post-inoculation talk and its role in building, strengthening, and spreading intra- and interindividual attitudinal resistance. One ripple at a time.

Chapter 5: Harnessing post-inoculation talk to confer intra- and interindividual resistance to persuasion

# GENERAL DISCUSSION

#### Chapter 6: General Discussion

This doctoral thesis set out to explore how inoculation theory may be leveraged to build, strengthen, and spread attitudinal resistance against misinformation. To do so, it revised and challenged the theoretical pillars threat and counterarguing and explored the role of attitude certainty and its effect on sharing misinformation. In essence, this thesis points towards theoretical mechanisms which, when conceptualised or operationalised differently, can shed light on and enhance the largely underexplored mechanisms underpinning the inoculation effect. Importantly, these mechanisms were explored through different interventions<sup>37</sup> (*Bad News, Join this Group, Go Viral, Tweet-sized PIT*), tested across varying misinformation contexts (e.g., vigilante riots, COVID-19, health misinformation), different platforms (e.g., simulated environments of Twitter, WhatsApp), and both within and between individuals (intra- and interindividual resistance to persuasion). In doing so, this thesis offers strong support for active, generalised, and therapeutic inoculation interventions against the spread of harmful misinformation and proposes a novel yet promising pathway towards achieving psychological herd immunity.

To briefly summarise, Chapter 2 offers support for generalised and gamified inoculation treatments against a series of common online misinformation techniques. More specifically, the findings suggest that compared to the control group, the active, generalised, and therapeutic inoculation intervention did not only successfully confer resistance but also boosted inoculated individuals' confidence in their ability to resist misinformation. Importantly, this confidence only occurred in the correct direction, meaning that it enhanced inoculated individuals' ability to confidently and correctly spot and resist misinformation techniques. Next, Chapter 3 further tested these findings with a national representative data set and extended the established efficacy of gamified inoculation treatments to the context of endto-end encrypted messaging applications where previously neglected psychological factors, such as group dynamics and peer pressure, are at force (Aizenkot & Kashy-Rosenbaum, 2018). Additionally, this chapter also shed initial light on attitude certainty as a potential mediator on sharing intentions of fake news items. In other words, the more certain individuals were in their reliability assessments, the less they shared misinformation. Crucially, this effect persisted over time, centring attitude certainty as a pre-requisite to effective build and strengthen resistance while also curbing the spread of misinformation. Though these findings are promising, they failed to account for a few critical factors: the reliance on previously seen items in the follow-up study, the length of the intervention, and the

<sup>&</sup>lt;sup>37</sup> Apart from *Bad News*, I contributed to the design, development, testing, and launch of all other gamified interventions reported in this doctoral thesis.

control condition's reliance on Tetris and thereby, inability to draw conclusions on whether active inoculation treatments are indeed more effective than passive treatments. Similar to the tendency of general inoculation research to assume theoretical factors, , recent studies on active inoculation have yet to identify the mechanistic underpinnings that facilitate the conferring of resistance.

Thus, Chapter 4 begins to disentangle the inoculation-based components within the gamified interventions. In other words, where in this choice-based game environment do threat and counterarguing occur and how, if at all, do these factors confer resistance to misinformation? Hence, Chapter 4 addresses these shortfalls and open questions by examining how the uptake, completion, and scalability of the "psychological vaccine" can be optimised. This Chapter reviews the efficacy of the 5-minute-long game developed in response to the pandemic (*Go Viral!*) against a passive prebunking intervention (UNESCO's pre-bunking infographics), with previously unseen items, and within a context that was changing and evolving alongside its launch. Importantly, the launch of *Go Viral!* presents a multi-disciplinary and real-world effort to spread an inoculation treatment. Specifically, in collaboration with the UK Cabinet Office and the World Health Organization (WHO), *Go Viral!* was part of an anti-misinformation campaign that reached millions of individuals across all continents since its launch. Thus, this chapter presents a substantial effort to spread inoculation-based resistance from one individual to another and, in the midst of an 'infodemic' explores one pathway towards curbing the spread of misinformation while simultaneously, encouraging the spread of inoculation.

Particularly, while the first two interventions inoculated against fictitious content that implemented manipulation techniques, Chapter 4 demonstrated the efficacy of inoculating against a topic which, at the time of its launch, was discussed world-wide, continuously updated by scientists, and regularly targeted by conspiratorial thinking (Cinelli et al., 2020; Desta & Mulugeta, 2020; Shahi et al., 2021). Contrary to the majority of inoculation research, this research was pre-registered and features open data and, across a multi-country RCT (with a partial nationally representative sample), provides strong support for the effectiveness of short and scalable prebunking interventions to reduce susceptibility to COVID-19 misinformation. Moreover, by adapting the recently suggested reconceptualisation of threat as a motivation rather than an apprehension-based concept (Banas & Richards, 2017), Chapter 4 demonstrated the predictive power of *motivational* threat in resistance against misinformation and thereby, consolidated its relevance in active, generalised, and therapeutic inoculation interventions. Furthermore, the findings further emphasise the role of attitude certainty in the ability to confidently spot manipulative content and withhold from sharing it with others online. On top of preventing the spread of misinformation, the first empirical comparison between passive and active inoculation treatments presented here, demonstrates that actively inoculated individuals also spread the inoculation treatment more. This finding sheds light on one possible pathway towards psychological herd immunity through inoculation.

#### Chapter 6: General Discussion

In a similar vein, Chapter 5 challenges the traditional conceptualisation of counterarguing, the theoretical counterpart to threat and, across three studies, questions whether the predominant focus on counterarguing as a subvocal process and demonstrates that, instead, PIT plays a role in conferring resistance, strengthening attitudes, bolstering attitude certainty, and spreading inter-individual resistance to persuasion through vicarious inoculation. Specifically, this chapter demonstrates that, post-inoculation talk (PIT) occurs organically and voluntarily. In other words, contrary to previous and limited research on PIT (Ivanov et al., 2016), inoculated individuals voluntarily engaged in talk after treatment exposure. By conceptualising PIT beyond the traditionally narrow lens on frequency and number of conversations, this chapter demonstrates that inoculated individuals feel more motivationally threatened and able to resist and refute manipulative content. Importantly, when asked what they would hypothetically pass on to others, inoculated individuals' intentional pass-along messages composed shared more general resistance and resistance-congruent content. The second phase demonstrates how *actual* talk that occurred in the interim between treatment exposure and the follow-up, strengthened attitudes, bolstered attitude certainty in inoculated individuals, and spread more resistance-congruent content than the control condition. Additionally, comparisons between intended and actual post-treatment talk demonstrated that inoculated individuals spread significantly more resistance in their *actual* talk than in their *intended* post-inoculation talk. Lastly, by exploring the efficacy of PIT messages in form of a replacement of traditional inoculation treatments, phase 3 provides initial and promising results for using inoculated individuals' talk as a mean to vicariously inoculate the recipients of talk. The effective passing-on of resistance between individuals, in turn, seems to trigger a "wave of resistance" where the vicariously inoculated also spread more inoculation and less misinformation in their respective pass-along messages. In establishing these effects, Chapter 5 carefully suggests an updated portrayal and use of PIT. More specifically, by first treating it as a form of vocalised counterarguing, then a process of inoculation, and lastly, as a product of inoculation, this chapter points towards a path in which PIT can be effectively leveraged toward psychological herd-immunity against online misinformation. Collectively, these Chapters shed light on the underexplored and predominantly assumed mechanisms underpinning the inoculation process. In critically assessing the theoretical key components of threat and counterarguing as well as the notion of what constitutes as resistance, this thesis further emphasises the need to rethink, update, and advance certain aspects of inoculation theory accordingly.

To summarise, despite the large body of inoculation research, a relatively small pool of studies has thus far addressed the mechanistic questions of how or why inoculation treatments are effective (Banas & Rains, 2010; Ivanov et al., 2018). In light of that, the present doctoral research aimed to expose the processes underpinning attitudinal resistance by focusing on the role of confidence, motivational threat, counterarguing and utilising vocalised forms of counterarguing as second-order inoculation treatments that can vicariously inoculate others. Next, I will discuss what these findings mean for the theoretical and practical application of inoculation.

# Attitude certainty

In *Paradise Lost*, Milton (1667) powerfully demonstrates the dangers of hubris – that is, an overconfidence accompanied by arrogance and a lack of humility, by describing the spiritual descend of Lucifer into Satan. Greek mythology provides another compelling symbol of hubris with the tale of Icarus who recklessly and overconfidently 'flew too close to the sun'. In the context of misinformation, research too points towards a pattern of societal overconfidence, highlighting a stark discrepancy between individuals' perceived ability to spot and resist misinformation and their actual ability to do so (Lyons et al., 2021). This introduces the element of ill-calibrated overconfidence and a "everyone-but-me" disposition of diffusing responsibility (Chung & Kim, 2021; Corbu et al., 2020). Put differently, individuals tend to believe that it is others that fall for and spread misinformation and demonstrate denialism when it comes to even considering that they might have unknowingly come across, believed in, and perhaps even shared misinformation with others before. Shaking this overconfidence and simultaneously operationalising it as a means to confer well-calibrated attitudinal resistance through inoculation theory constituted a crucial focus of this thesis.

Though general persuasion research emphasised the role of attitude certainty in resisting persuasion, strengthening attitudes, and advocating for one's beliefs (Brügger & Höchli, 2019; Tormala & Petty, 2004), inoculation research, arguably the most prominent account of empirically studying resistance to persuasion, has given it little attention. As a consequence, the role of attitude certainty in the resistance process remains opaque, leading some scholars to call for a more meta-cognitive approach to the study of resistance Tormala & Petty, 2002). Indeed, the authors question the notion that successful resistance stops at no attitudinal change and, instead, argue that research needs to address whether one's meta-cognitive awareness of the process of resistance seems to impact how confidently attitudes are held (Tormala & Petty, 2007). This suggests that attitude certainty may play a crucial role in building and strengthening attitudinal resistance to persuasion, a notion that had scarcely been studied by inoculation research and had yet to be applied to the context of active, generalised, and therapeutic inoculation interventions.

Building on that, the findings presented in this thesis (specifically, Chapters 2, 3, and 4) suggest that attitude certainty plays a crucial role in the inoculation process. Specifically, enhanced confidence in *correct* reliability assessments and its mediating effect on sharing intentions of misinformation indicate that a meta-cognitive awareness of one's certainty impacts the generation and strengthening of inoculation-based resistance to persuasion. Importantly, and contrary to the pitfalls of hubris mentioned above, the presented research illustrates a *correct* confidence boost post-treatment exposure. In other words, only when the post-treatment reliability assessments of misinformation indicated individuals get more confident in their ability to resist. Furthermore, the mediating role of attitude certainty on sharing intentions of fake news items emphasise the need for inoculation interventions to prioritise the properties which correctly boost confidence to prevent

misinformation from being shared with in the first place. However, since the study designs did not allow to test topic-specific factual knowledge nor gain insights into inoculated individuals' actual sharing behaviour on their social media accounts, future research should set out to explore just *how* well-calibrated inoculation-induced confidence and resistance is in the real world (Fischer et al., 2019; Hinsley & Holton, 2021).

# Updating the fundamentals: threat, counterarguing

For attitude certainty to be used to enhance and optimise inoculation-based resistance, the theoretical pillars of inoculation theory, threat and counterarguing, needed to be first reviewed. Specifically, though both assumed and treated as critical components to resistance through inoculation, a metaanalysis revealed how threat did, in fact, not significantly predict resistance (Banas & Rains, 2010). Accordingly, Chapter 4 aimed to disentangle the operationalisation of threat and address the chasm between theoretical assumption and empirical reality in the inoculation scholarship. More specifically, while early theorising describes threat as a motivation and a catalyst which triggers the attitudebolstering generation of counterarguments (McGuire, 1964; Compton & Pfau, 2005), scholars have pointed out how the traditional operationalisation of threat is based on of anxiety, fear, and apprehension (Banas & Richards, 2017). In short, one that seems to be at odds with the descriptively motivating nature of a threat.

By incorporating a comparison of both approaches to threat to an unprecedented and uncertain context such as the pandemic, Chapter 4 found promising support for the case of motivational threat. Indeed, the results suggest that individuals inoculated through *Go Viral!* experienced significantly higher levels of motivational threat (not apprehensive threat). The results further suggest that only motivational threat predicted the perceived manipulativeness of misinformation items. Thus, Chapter 4 is, to my knowledge, the first study to demonstrate the superiority of motivational threat in predicting resistance within the context of active, generalised, and therapeutic inoculation treatments. Considering that *Go Viral!* was launched and tested prior to any vaccinations being available, the finding that inoculated individuals feel more *motivated* to equip themselves with the necessary skills to resist misinformation in a highly uncertain and risky context such as the pandemic, offers strong support for Banas and Richard's (2017) proposition to revisit and rethink threat.

Similarly, inoculation theory positions counterarguing as the second building block to conferring attitudinal resistance. However, much of inoculation research so far has limited the process of counterarguing to an exclusively subvocal conceptualisation (Compton & Pfau, 2009; Ivanov et al., 2012). In other words, counterarguing is traditionally approached and measures as an intrapersonal dialogue with the anticipated attack message – a process in which the individual internally raises and refutes arguments against the anticipated attack message. Thus, research exploring subvocal as well as vocalised counterarguing (i.e., talk) has been limited (Ivanov et al., 2016). Consequently, Chapter 5 set out to address the current gaps in our understanding of vocalised counterarguing, or post-

inoculation talk (PIT), and how, in turn, it might be utilised to spread attitudinal resistance from one person to another. While the previous chapters demonstrated that inoculated individuals reported significantly lower intentions to share misinformation, this chapter aimed to take it one step further and examine whether harnessing PIT could lead inoculated individuals to share more inoculation.

In a novel three-part study design, which employed mixed research methods, the results provide multiple substantial insights. First, inoculated individuals do voluntarily engage in post-inoculation talk and doing so strengthens their attitudes, attitude certainty, and attitude strength. Next, contentbased analyses illustrate how, compared to the control condition, inoculated individuals spread significantly more inoculation-congruent material (e.g., forewarning others, pre-emptively debunking manipulation techniques or credibility of chemical) in their intended as well as actual talk when recalled a week later. Lastly, when individuals receive second-order inoculation treatments, that is, when post-inoculation talk is used as a treatment and is pitted against misinformation, attitudinal resistance is conferred *vicariously*. Considering that the pass-along messages of the vicariously inoculated individuals also spread less misinformation and more inoculation, these findings highlight the possibility of post-inoculation talk serving as a pathway towards psychological herd immunity. Although the focus on high-quality PIT, that is strong messages shared with a large number of individuals (i.e., "waves of resistance"), is one way to pursue psychological herd immunity, future research should also investigate the alternate route of "the ripple effect", that is, one post-inoculation message travelling through various chains of communication. Or put simply, how many times can one post-inoculation message be passed on until it loses its effectiveness in conferring resistance to its recipients? Lastly, just as research has yet to establish how frequency and intensity of misinformation exposure leads to persuasion (van der Linden, 2022), future research should aim to identify the prerequisites for PIT to be effective in an environment in which individuals make news-sharing decisions.

# Resistance in the digital age

Importantly, albeit not intentionally, Chapter 5 shed critical light on the conceptualisation of resistance in contemporary inoculation research. To begin with, most research on resistance to persuasion assumed that no attitude change equalled attitudinal resistance (Ivanov et al., 2012). Not only does this approach disregard the role of attitudinal ascendents, such attitude certainty, accessibility, and strength in the resistance process, this theoretical assumption is also ill-fitted for today's information infrastructure. Decades before the advent of the internet, the artist Andy Warhol (1968), famously said "In the future, everyone will be world-famous for 15 minutes.". This idea and now, reality of a fastpaced media environment where information travels faster and further than before and can lead to the rapid formation and collapse of 'collective thoughts" is perhaps best demonstrated through the phenomenon of viral altruism (van der Linden, 2017). To put it into the context of inoculation theory, one might therefore question whether, in an environment where individuals can rapidly form and abandon opinions and beliefs, is the absence of attitude change still the most reliable means to capture resistance? Or, as suggested in Chapter 5, is it possible that a generalised and therapeutic inoculation treatment can only keep up with or even outpace misinformation when it focuses making people, regardless of their position, better at spotting, resisting, and refraining from passing on misinformation when they encounter it online? After all, Chapter 5 shows how, despite neither the inoculation nor control condition experiencing any attitude changes toward the non-existent, fictitious chemical, inoculated individuals perceived their ability to spot and resist misinformation to be higher. And perhaps more importantly, inoculated individuals, in both their intended and actual post-treatment messages, shared less misinformation and more inoculation with others – a pattern that was repeated by the vicariously inoculated. This suggests that tectonic shifts in attitudes (or their lack of) do not paint a sufficiently nuanced picture of individuals' actual ability to resist in the digital age.

## The case for active, generalised, and therapeutic inoculations

In sum, the Chapters 2, 3, and 4 provide support for novel inoculation interventions against misinformation. In doing so, these chapters emphasise the efficacy of therapeutic inoculation treatments on the "already afflicted" (Wood, 2007) as well as against societally contested issues (van der Linden et al., 2017). In other words, by demonstrating the efficacy of different gamified inoculation treatments across longitudinal, cross-cultural, and nationally representative samples, this thesis strongly supports the efficacy of therapeutic inoculation treatments. Furthermore, its application to varying platforms and contexts further highlights that inoculation is not contextually bound. Compton (2013) argues that the notion of therapeutic inoculation against differing attitudes is not necessarily inconsistent with the analogic but rather, holds tightly to it. Instead, similar to the varying incubation periods of viruses, therapeutic vaccines can be designed to simultaneously treat an existing illness while also providing immunisation against future attacks (Nossal, 1999; Compton et al., 2020). Importantly, therapeutic inoculation messages applied to individuals with differing positions on the issue appear to avoid psychological reactance and backfire effects (Jolley & Douglas, 2017; van der Linden et al., 2017; Williams & Bond, 2020).

Alongside the avoidance of unwanted consequences, Chapter 2 and 3 further highlight how therapeutic inoculation can extend its "blanket of protection" to misinformation techniques applied to unmentioned arguments within the same issue topic. To take it a step further, Chapter 4 and 5 aid to the currently underexplored and somewhat misunderstood concept of the 'spill-over effect' of attitudinal resistance and aimed to disentangle inoculation scholarships' opaque understanding of and interchangeable use of terminologies describing different phenomena of spreading resistance. Specifically, Chapter 4 demonstrated inoculation's ability to confer "cross-protection", that is enhance inoculated individuals' ability to spot and resist contextually related yet previously unseen forms of manipulation (i.e., new set of fake news items on different arguments ranging from adherence to health guidelines and vaccine hesitancy to silencing scientific findings). Building on that, Chapter 5 provides evidence for an "umbrella of protection", where the inoculation message does cover the same manipulation techniques underpinning the attack message yet apply them to *several* distinctly different

topics that are unrelated ant untreated to the focus of the attack message (a fictitious chemical). Even more so, considering that the arguments in the attack message focus on conceptually different risks (e.g., chemicals not being vigorously tested before allowed on the market) compared to the manipulation-same arguments (here, conspiracy theories) prebunked in the inoculation treatment (e.g., scientists ignoring doctors' concerns about 5G). Thus, while the content shared the overarching issue health, the inoculation treatment effectively inoculated against the same manipulation strategy on different topics. To take it a step further, the efficacy of the treatment in light of the distinctly different topics within the treatment message itself (e.g., 5G radiation, ear buds, toothpaste) provides further support for inoculation conferring an "umbrella of protection" against health misinformation. In other words, Chapter 5 suggests that traditional, or rather passive, text-based inoculation treatments do not need to consist of a concentrated set of refutations on one topic for it to trigger the production of "mental antibodies" and to confer a "blanket of protection" against an untreated and unrelated topic using the same manipulation techniques on *different* arguments.

# Where do we go from here?

Indeed, Compton (2013) suggests that the biological analogy of inoculation theory should serve as a mere compass rather than a constraining prescription. In challenging the boundary conditions of therapeutic inoculation interventions across a series of studies, this doctoral thesis provides strong support for inoculating those who may or not have come in contact with the 'virus' of misinformation about differing, societally contested, and often incompletely researched and scientifically understood topics. To take it a step further, I would like to challenge contemporary inoculation scholarship to push past the presumed theoretical and practical boundaries of inoculation and rather, use the biological analogy as a 'springboard' that can inspire but not limit the ways in which we attempt to build, strengthen, and spread intra- and interindividual resistance against one of the most critical societal challenges we face today. This is an exciting time for inoculation research, despite its "theoretical maturity" (Compton & Pfau, 2005), it has experienced renewed interest which, in turn, led to novel and substantial innovations, impacting the way in which we think about and work towards resistance to persuasion. But just as a virus is constantly evolving, inoculation research needs to keep up with the "mutations and variants" of misinformation just as much as it must keep up with the fast-paced and ever-changing environments in which communication takes place today. Significant changes occurred since van der Linden and colleagues (2017) first applied inoculation theory to the context of misinformation about climate change. However, it is somewhat perplexing that inoculation theory has predominantly maintained an overtly and almost exclusively cognitive account of resistance to persuasion. Misinformation and its more malign forms (i.e., disinformation, propaganda), have adapted to and made use of the 'breeding grounds' of social media (Mandical et al., 2020). Which is why I would argue that in the current Zeitgeist of explicit deep fakes, reels and TikTok videos, elaborative podcast episodes, trending hashtags, and increasingly polarising communities tucked away in end-to-end encrypted group chats, not incorporating the role of emotion, morality, and the

complexity of today's communication infrastructure into inoculation scholarship will, though regarded as the grandfather theory of persuasion, debilitate its ability to stay relevant against the serious threats posed by misinformation.

In similar vein, three key limitations present in this doctoral thesis point towards future avenues for inoculation research in the digital age. Firstly, although the efficacy of gamified and therapeutic inoculation treatments is demonstrated across multiple chapters, research on it thus not yet allow to determine whether these games (Bad News, Join this Group, and Go Viral!) lead to an enhanced ability to spot and resist misinformation in their actual online behaviour (i.e., on their personal social media accounts). While this was primarily restricted by data protection concerns, future research should examine whether there are appropriate ways to establish the inoculation effect on resistance against misinformation in the real-world. Furthermore, although this is also true for the other gamified interventions discussed in this thesis, Join this Group and its particular effort to examine the efficacy of inoculation within a context that is specific to private group chats, suffered, despite its interface simulation, from the fact that the messages were not coming from one's own contacts. Considering research that emphasises the effects of perceived source credibility on (resistance to) persuasion (Traberg & van der Linden, 2022), future research must examine the role of sources (e.g., one's own social circle) when inoculating against misinformation spread on private messaging platforms. Similarly, post-inoculation talk, as a potential pathway to spread resistance between individuals, was explored in a one-directional way. While this allowed to first establish whether PIT takes place, effects the spreader, and can be passed on, future research is needed to situate PIT in the dynamics of a conversation. In other words, more research is needed to understand how endorsement and opposition encountered in a dialogue might affect the inoculation process for both the spreader and the recipient of vicarious treatments.

# Lessons learned and take-away messages

Across disciplines, many appear to adopt the 'loom and gloom' narrative of misinformation (Effron & Helgason, 2022; Iyengar & Massey, 2019; Runciman, 2016). On the contrary, I wish to advocate for a more optimistic stance on the future of our information landscape and psychology's part in it. Throughout this doctoral thesis, I have presented highly collaborative and applied work that situates inoculation at the centre of the ongoing fight against misinformation. Reflections on these collaborations point to multiple challenges and benefits.

First, our work with WhatsApp was part of a larger effort to fight the spread of misinformation on their platform with a particular focus on India. At the time of the project, misinformation in India undermined political elections and fuelled vigilante groupings, violence, and mob lynchings (Arun, 2019; Kazemi et al., 2021). Developing *Join this Group* thus, represented an effort to translate the well-established efficacy of gamified, active, and therapeutic inoculation treatments into a culturally

different context. However, we failed to replicate the findings in the Indian context (Harjani, Basol, et al., forthcoming). This sheds light on a larger issue evident across the inoculation scholarship but also our psychological discipline as a whole. Consequently, any attempt to stop the global spread of misinformation, will only be effective when it moves away from the WEIRDness of our growing scientific understanding and starts addressing the cultural differences and varying societal structures underpinning susceptibility to misinformation (Adams & Estrada-Villalta, 2017; Hansen & Heu, 2020; Henrich et al., 2010). Thus, our ongoing collaboration with WhatsApp has highlighted the importance to translate and adequately contextualise interventions to the respective cultures. For instance, during an ongoing project with WhatsApp, we noted stark cultural discrepancies between the UK and India in regards to how the platform is used (Hanraji, Basol, et al., forthcoming). For instance, Indian users report using the app to retrieve news and health and safety related information, being part of significantly larger group chats that are not limited to phone contacts, reporting higher levels of consciously spreading misinformation, mistrusting the government, and seeing it as their responsibility to enforce law and order. This is consistent with research emphasising that a large portion of India's misinformation campaigns are directed at minority communities - Dalits and Muslims, emphasising the unchartered territory of religious ideology, media habits, and violent behaviour fuelled by misinformation (Mukherjee, 2020). Thus, I urge future inoculation research to prioritise replicating the wealth of evidence on "the grandparent theory of resistance to attitude change" (Eagly & Chaiken, 1993, p.561) beyond the Global North.

In that vein, in our collaboration with the UK Cabinet Office for *Go Viral!*, we prioritised the spread and effectiveness of an inoculation intervention. Indeed, at the time of this write-up, the freely accessible gamified inoculation intervention is available in 13 languages (including Ukrainian, Russian, and Brazilian Portuguese) and was played by over 1.4 million individuals across the world. By joining forces with the UN and WHO, this project launched a cross-cultural inoculation campaign against COVID-19 misinformation and conspiracies and highlighted the applicability, scalability, and impact that psychologically informed interventions can have in the real world<sup>38</sup>. Of course, this project too came with challenges and the respective lessons that emerged from them. Firstly, the empirical comparison of *Go Viral!*, a generalised, active, and therapeutic inoculation intervention with UNESCO's infographics, a tool, which alongside other text-based efforts like brochures and posters,

<sup>&</sup>lt;sup>38</sup> <u>https://en.unesco.org/news/qa-inoculating-against-covid-19-misinformation</u>

represents a more traditional form of campaigning. While these specific infographics where informed by inoculation scholars and therefore, prebunked misinformation in their framing, research suggests that poorly designed infographics can distort viewers' understanding and decrease the perceived credibility of the information (Freeman et al., 2021; Parrott et al., 2005). Critically assessing what constitutes effective science communication throughout this project, has emphasised the need for interventions that employ adequate verbal and visual simplicity to render itself accessible and trustworthy to individuals ranging in demographics (e.g., age, education, disabilities) as well as levels of health, media, and digital literacy (Jalil et al., 2021; Rasi et al., 2019; U.S. Department of and Education, 2017). Overall, this collaboration highlighted the room for improvement when it comes to incorporating evidence-based interventions into the policy-making process and illustrated the friction that can result from inter-disciplinary differences in focus points and decisions (e.g., preferred framing, depth, and length of intervention). This allowed me to practice the careful balancing act of objecting as an academic advisor (e.g., not negotiating the necessity to test and establish efficacy of an intervention in a well-powered randomised controlled trial prior to its launch) while responding to time-sensitive challenges in the midst of the pandemic. Simultaneously, Go Viral! was an opportune project to challenge and advance the boundary conditions of inoculation theory by establishing the minimum criteria necessary for inoculation interventions to effectively confer attitudinal resistance.

To summarise, these projects reflect unique opportunities to be part of the governmental and institutional problem-solving process at the hight of crises, ranging from misinformation undermining political elections to complicating the mitigation of the pandemic. As a result, I hope that both the theoretical and practical work presented throughout this thesis assists the implementation of future inoculation interventions. Firstly, I hope that the theoretical advancements of this work provide a foundational building block to ground, inspire, and guide future inoculation researchers committed to applying their work to the societally pressing issues we face today. Whether that concerns conspiratorial thinking and vaccine hesitancy during a global health crisis or conferring resistance against Russian propaganda (Alyukov, 2022; Desta & Mulugeta, 2020), the need for more applied inoculation research to real-world issues is undeniable. Secondly, I hope that this thesis adequately highlighted that harnessing 'the psychology of fake news'' as well as a *proactive* and *pre-emptive* commitment to stop misinformation at the starting line are critical components to succeeding against the pervasiveness of misinformation.

However, I would also like to emphasise that inoculation theory offers no "silver bullet" against misinformation and that, and perhaps this is a greater issue within our discipline as a whole, we must fight the temptation to make overarching and overselling promises when communicating and applying our research to the real world (Chater & Loewenstein, 2022). Instead, a multi-layered defence mechanism incorporating legislative, technological, and educational efforts is needed to effectively combat the complex and deep-rooted symptoms *and* causes of misinformation (Compton et al., 20201). To do so, I call for scholars working on misinformation to take a step back and critically

confront the status quo of how misinformation is studies today. A multi-pronged approach against misinformation will only be effective if, instead of disproportionately focusing on the individual, it addresses the underlying societal structures and deeper-rooted causes underlying prevalence of misinformation. In fact, scholars have criticised the analogy of an "infodemic", arguing that it simplifies the spread of misinformation which, contrary to a virus, involves unclear origins, boundaries, and interpretations and fails to account for the complex psycho-social underpinnings (Simon & Camargo, 2021). Whether in form of digital divides (i.e., information deserts, disproportionate access to technology, the internet, and media literacy) or culturally engrained, and often, valid scepticism towards experts, scientists, and governments, an overreliance on technological interventions risks fighting the symptoms rather than the causes misinformation (Champion, 2009; Sparks, 2014). Thus, I urge future researchers and decision-makers to think about how to reach those that may rely on more traditional means of accessing information (e.g., local newspaper, word-of-mouth communication) and may not be as receptive to the inoculation interventions tested thus far. A societal challenge as vast as the prevalence of misinformation will require collaborative, evidence-based, and actionable steps. I hope that doctoral thesis takes one such step into that direction.

# 7 References

Abbott, J. (2013). Introduction: Assessing the Social and Political Impact of the Internet and New Social Media in Asia. *Journal of Contemporary Asia*, 43(4), 579–590. https://doi.org/10.1080/00472336.2013.785698

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, 5(1), 1–7.

Adams, G., & Estrada-Villalta, S. (2017). Theory from the South: a decolonial approach to the psychology of global inequality. *Current Opinion in Psychology*, *18*, 37–42. https://doi.org/10.1016/J.COPSYC.2017.07.031

Adriani, R. (2019). *The Evolution of Fake News and the Abuse of Emerging Technologies*. 2(1). https://doi.org/10.26417/ejss.v2i1.p32-38

Aizenkot, D., & Kashy-Rosenbaum, G. (2018). Cyberbullying in WhatsApp classmates' groups: Evaluation of an intervention program implemented in Israeli elementary and middle schools. *New Media & Society*, *20*(12), 4709–4727. https://doi.org/10.1177/1461444818782702

Akhtar, O., Paunesku, D., & Tormala, Z. L. (2013). Weak > strong: the ironic effect of argument strength on supportive advocacy. *Personality & Social Psychology Bulletin*, *39*(9), 1214–1226. https://doi.org/10.1177/0146167213492430

Albarracín, D., & Mitchell, A. L. (2004). The role of defensive confidence in preference for proattitudinal information: how believing that one is strong can sometimes be a defensive weakness. *Personality & Social Psychology Bulletin, 30*(12), 1565–1584. https://doi.org/10.1177/0146167204271180

Albarracín, D., Wyer, R. S., & Jr. (2000). The cognitive impact of past behavior: influences on beliefs, attitudes, and future behavioral decisions. *Journal of Personality and Social Psychology*, 79(1), 5–22. http://www.ncbi.nlm.nih.gov/pubmed/10909874

 Allen, M. (2009). Meta-analysis comparing the persuasiveness of one-sided and two-sided messages.

 *Http://Dx.Doi.Org/10.1080/10570319109374395*,
 55(4),
 390–404.

 https://doi.org/10.1080/10570319109374395
 55(4),
 390–404.

Alyukov, M. (2022). Propaganda, authoritarianism and Russia's invasion of Ukraine. *Nature Human Behaviour 2022 6:6*, *6*(6), 763–765. https://doi.org/10.1038/s41562-022-01375-x

Al-Zaman, M. S. (2021). A Thematic Analysis of Misinformation in India during the COVID-19Pandemic.InternationalInformationandLibraryReview.https://doi.org/10.1080/10572317.2021.1908063

Anderson, L. R. (1967). Belief defense produced by derogation of message source. *Journal of Experimental Social Psychology*, 3(4), 349–360. https://doi.org/10.1016/0022-1031(67)90003-0

Angelopoulos, C., Brody, A., Hins, W., Hugenholtz, B., Leerssen, P., Margoni, T., Mcgonagle, T., van Daalen, O., & van Hoboken, J. (2016). *Study of fundamental rights limitations for online enforcement through self-regulation.* www.ivir.nl

Angelopoulos, C., & Smet, S. (2016). Notice-and-fair-balance: how to reach a compromise betweenfundamentalrightsinEuropeanintermediaryliability.*Https://Doi.Org/10.1080/17577632.2016.1240957*,8(2),266–301.https://doi.org/10.1080/17577632.2016.1240957

Arun, C. (2019). On WhatsApp, Rumours, Lynchings, and the Indian Government. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3336127

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

Ball, P., & Maxmen, A. (2020). The epic battle against coronavirus misinformation and conspiracy theories. *Nature*, *581*(7809), 371–374. https://doi.org/10.1038/D41586-020-01452-Z

Ballew, M. T., Goldberg, M. H., Rosenthal, S. A., Gustafson, A., & Leiserowitz, A. (2019). Systems thinking as a pathway to global warming beliefs and attitudes through an ecological worldview. *Proceedings of the National Academy of Sciences*, *116*(17), 8214–8219. https://doi.org/10.1073/pnas.1819310116

Banaji, S., With, R. B., Agarwal, A., Passanha, N., & Pravin, M. S. (2019a). WhatsApp Vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India.

https://eprints.lse.ac.uk/104316/1/Banaji\_whatsapp\_vigilantes\_exploration\_of\_citizen\_reception\_pu blished.pdf

Banaji, S., With, R. B., Agarwal, A., Passanha, N., & Pravin, M. S. (2019b). WhatsApp Vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. https://www.lse.ac.uk/media-andcommunications/assets/documents/research/projects/WhatsApp-Misinformation-Report.pdf

Banas, J. A., & Miller, G. (2013). Inducing Resistance to Conspiracy Theory Propaganda: Testing Inoculation and Metainoculation Strategies. *Human Communication Research*, *39*(2), 184–207. https://doi.org/10.1111/hcre.12000

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

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

Banas, J. A., & Richards, A. S. (2017a). Apprehension or motivation to defend attitudes? exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. *Communication Monographs*, 84(2), 164–178.

Banas, J. A., & Richards, A. S. (2017b). Communication Monographs Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. https://doi.org/10.1080/03637751.2017.1307999

Barden, J., & Petty, R. E. (2008). The mere perception of elaboration creates attitude certainty: Exploring the thoughtfulness heuristic. *Journal of Personality and Social Psychology*, 95(3), 489–509. https://doi.org/10.1037/a0012559

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. van der. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, 8(1), 205395172110138. https://doi.org/10.1177/20539517211013868

Basol, M., Roozenbeek, J., & van der Linden, S. (2020a). 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

Basol, M., Roozenbeek, J., & van der Linden, S. (2020b). Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition*, *3*(1), 2. https://doi.org/10.5334/joc.91

Bassili, J. (1996). Meta-judgmental versus operative indexes of psychological attributes: The case of measures of attitude strength. *Journal of Personality and Social Psychology*, *71*(4), 637–653.

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *17*(1), 78–98. https://doi.org/10.1177/1745691620986135 BBC News. (2020, March 8). Coronavirus: The fake health advice you should ignore. *Www.Bbc.Co.Uk*.

Belli, L., & Cavalli, O. (2019). Law of the Land or Law of the Platform? Beware of the Privatisation of Regulation and Police. In *Internet governance and regulations in Latin America : analysis of infrastructure, privacy, cybersecurity and technological developments in honor of the tenth anniversary of the South School on Internet Governance.* www.fgv.br/direitorio

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions: *Https://Doi.Org/10.1177/0267323118760317*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bernard, M. M., Maio, G. R., & Olson, J. M. (2003). The vulnerability of values to attack: Inoculation of values and value-relevant attitudes. *Personality and Social Psychology Bulletin*, 29(1), 63–75. https://doi.org/10.1177/0146167202238372

Berriche, M., & Altay, S. (2020). Internet users engage more with phatic posts than with health misinformation on Facebook. *Palgrave Communications*, 6(1), 1–9.

Bertolini, A., & Aiello, G. (2018). Robot companions: A legal and ethical analysis. *The Information Society*, *34*(3), 130–140. https://doi.org/10.1080/01972243.2018.1444249

Bleakley, P. (2021). Panic, pizza and mainstreaming the alt-right: A social media analysis of Pizzagate and the rise of the QAnon conspiracy. *Current Sociology*, 001139212110348. https://doi.org/10.1177/00113921211034896

Bleize, D. N. M., Tanis, M., Anschütz, D. J., & Buijzen, M. (2021a). A social identity perspective on conformity to cyber aggression among early adolescents on WhatsApp. *Social Development*, sode.12511. https://doi.org/10.1111/sode.12511

Bleize, D. N. M., Tanis, M., Anschütz, D. J., & Buijzen, M. (2021b). A social identity perspective on conformity to cyber aggression among early adolescents on WhatsApp. *Social Development*, sode.12511. https://doi.org/10.1111/sode.12511

Bonetto, E., Troïan, J., Varet, F., lo Monaco, G., & Girandola, F. (2018). Priming Resistance toPersuasiondecreasesadherencetoConspiracyTheories\*.*Https://Doi.Org/10.1080/15534510.2018.1471415*,*13*(3),125–136.https://doi.org/10.1080/15534510.2018.1471415III (International International Inte

Bordia, P., & DiFonzo, N. (2017). *Rumor Mills: the Social Impact of Rumor and Legend*. https://books.google.co.uk/books?id=hD4rDwAAQBAJ&dq=DiFonzo+2017+rumour+psychology& lr=&hl=de&source=gbs\_navlinks\_s

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, 28(3), 771–780. https://doi.org/10.1016/J.CHB.2011.11.020

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brannon, L. A., Tagler, M. J., & Eagly, A. H. (2007). The moderating role of attitude strength in selective exposure to information. *Journal of Experimental Social Psychology*, *43*(4), 611–617. https://doi.org/10.1016/J.JESP.2006.05.001

Brechwald, W. A., & Prinstein, M. J. (2011). Beyond Homophily: A Decade of Advances in Understanding Peer Influence Processes. *Journal of Research on Adolescence*, *21*(1), 166–179. https://doi.org/10.1111/j.1532-7795.2010.00721.x

Champion, W. T. Jr. (2009). The Tuskegee Syphilis Study as a Paradigm for Illegal, Racist, and Unethical Human Experimentation. *Southern University Law Review*, *37*. https://heinonline.org/HOL/Page?handle=hein.journals/soulr37&id=235&div=14&collection=journa ls

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11), 1531–1546. https://doi.org/10.1177/0956797617714579

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

Cheatham, L., & Tormala, Z. L. (2015). Attitude Certainty and Attitudinal Advocacy: The Unique Roles of Clarity and Correctness. *Personality and Social Psychology Bulletin*, *41*(11), 1537–1550. https://doi.org/10.1177/0146167215601406

Chung, M., & Kim, N. (2021). When I Learn the News is False: How Fact-Checking Information Stems the Spread of Fake News Via Third-Person Perception. *Human Communication Research*, 47(1), 1–24. https://doi.org/10.1093/HCR/HQAA010

Cialdini, R., & Trost, M. (1998). Social influence: Social norms, conformity and compliance. *Undefined.* https://www.semanticscholar.org/paper/Social-influence%3A-Social-norms%2C-conformity-and-Cialdini-Trost/bc4d09459f298901ebb6894652319c9be3c3b8b2

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

Compton, J. (2013a). Inoculation Theory. In *The SAGE Handbook of Persuasion: Developments in Theory and Practice* (pp. 220–236). SAGE Publications, Inc. https://doi.org/10.4135/9781452218410.n14

Compton, J. (2013b). 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, Inc. https://doi.org/10.4135/9781452218410

Compton, J. (2019). Prophylactic Versus Therapeutic Inoculation Treatments for Resistance to Influence. https://doi.org/10.1093/ct/qtz004

Compton, J., Jackson, B., & Dimmock, J. A. (2016). Persuading Others to Avoid Persuasion: Inoculation Theory and Resistant Health Attitudes. *Frontiers in Psychology*, 7, 122. https://doi.org/10.3389/fpsyg.2016.00122

Compton, J., 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

Compton, J., & Pfau, M. (2004). Use of inoculation to foster resistance to credit card marketing targeting college students. *Journal of Applied Communication Research*, *32*(4), 343–364. https://doi.org/10.1080/0090988042000276014

Compton, J., & Pfau, M. (2005a). Inoculation Theory of Resistance to Influence at Maturity: Recent Progress In Theory Development and Application and Suggestions for Future Research. *Annals of the International Communication Association*, 29(1), 97–145. https://doi.org/10.1207/s15567419cy2901 4

Compton, J., & Pfau, M. (2009). Spreading inoculation: Inoculation, resistance to influence, and word-of-mouth communication. *Communication Theory*, *19*(1), 9–28. https://doi.org/10.1111/j.1468-2885.2008.01330.x

Compton, Josh., & Pfau, M. (2005b). Inoculation Theory of Resistance to Influence at Maturity: Recent Progress In Theory Development and Application and Suggestions for Future Research. *Annals of the International Communication Association*, 29(1), 97–146. https://doi.org/10.1080/23808985.2005.11679045

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017a). 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

Corbu, N., Oprea, D. A., Negrea-Busuioc, E., & Radu, L. (2020). 'They can't fool me, but they can fool the others!' Third person effect and fake news detection: *Https://Doi.Org/10.1177/0267323120903686*, *35*(2), 165–180. https://doi.org/10.1177/0267323120903686

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

David, M. (2015). New Social Media: Modernisation and Democratisation in Russia. *European Politics and Society*, *16*(1), 95–110. https://doi.org/10.1080/15705854.2014.965892

de Freitas Melo, P., Vieira, C. C., Garimella, K., de Melo, P. O. S. V., & Benevenuto, F. (2020). *Can WhatsApp Counter Misinformation by Limiting Message Forwarding?* (pp. 372–384). Springer, Cham. https://doi.org/10.1007/978-3-030-36687-2\_31

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., & Larson, H. (2020). The pandemic of social media panic travels faster than the COVID-19 outbreak. In *Journal of travel medicine* (Vol. 27, Issue 3). https://doi.org/10.1093/jtm/taaa031

Desta, T. T., & Mulugeta, T. (2020). Living with COVID-19-triggered pseudoscience and conspiracies. *International Journal of Public Health* 2020 65:6, 65(6), 713–714. https://doi.org/10.1007/S00038-020-01412-4

DiFonzo, N., & Bordia, P. (2011). Rumors Influence. In *The Science of Social Influence* (pp. 271–295). Psychology Press. https://doi.org/10.4324/9780203818565-11

Dillingham, L. L., & Ivanov, B. (2016a). Using Postinoculation Talk to Strengthen GeneratedResistance.CommunicationResearchReports,33(4),295–302.https://doi.org/10.1080/08824096.2016.1224161

Dillingham, L. L., & Ivanov, B. (2016b). Using Postinoculation Talk to Strengthen GeneratedResistance.CommunicationResearchReports,33(4),295–302.https://doi.org/10.1080/08824096.2016.1224161

Dinauer, L. D., & Fink, E. L. (2005). Interattitude Structure and Attitude Dynamics. *Human Communication Research*, *31*(1), 1–32. https://doi.org/10.1111/j.1468-2958.2005.tb00863.x

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding Conspiracy Theories. *Political Psychology*, *40*(S1), 3–35. https://doi.org/10.1111/pops.12568 Dryhurst, S., Schneider, C. R., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., Spiegelhalter, D., & van der Linden, S. (2020). Risk perceptions of COVID-19 around the world. *Journal of Risk Research*, 23(7–8), 994–1006. https://doi.org/10.1080/13669877.2020.1758193

Eagly, A., & Chaiken, S. (1997). Attitude structure and function. In *The Handbook of Social Psychology* (pp. 269–322). Oxford University Press.

Eagly, A. H., & Chaiken, S. (1993). The Psychology of Attitudes. Harcourt Brace Jovanovich.

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., 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* 2022 1:1, 1(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., Lewandowsky, S., Jayawardana, K., & Mladenovic, A. (2019). Refutations of equivocal claims: No evidence for an ironic effect of counterargument number. *Journal of Applied Research in Memory and Cognition*, 8(1), 98–107. https://doi.org/10.1037/H0101834

Effron, D. A., & Helgason, B. A. (2022). The Moral Psychology of Misinformation: Why We Excuse Dishonesty in a Post-Truth World. *Current Opinion in Psychology*, 101375. https://doi.org/10.1016/J.COPSYC.2022.101375

Effron, D. A., & Raj, M. (2019). Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share. *Psychological Science*, *31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Effron, D. A., & Raj, M. (2020). Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share. *Psychological Science*, *31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Ellis-Peterse, H. (2021). *WhatsApp sues Indian government over 'mass surveillance' internet laws India The Guardian*. https://www.theguardian.com/world/2021/may/26/whatsapp-sues-indian-government-over-mass-surveillance-internet-laws

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

Fazio, L. K., 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

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*. https://doi.org/10.5210/fm.v25i6.10633
Fischer, H., Amelung, D., & Said, N. (2019). The accuracy of German citizens' confidence in their climate change knowledge. *Nature Climate Change* 2019 9:10, 9(10), 776–780. https://doi.org/10.1038/s41558-019-0563-0

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Fletcher, R., & Nielsen, R. K. (2018). Generalised scepticism: how people navigate news on socialmedia.InformationCommunicationandSociety,22(12),1–19.https://doi.org/10.1080/1369118X.2018.1450887

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. *Political Psychology*, *38*, 127–150. https://doi.org/10.1111/POPS.12394

Freedman, J. L., & Sears, D. O. (1965). Warning, distraction, and resistance to influence. *Journal of Personality and Social Psychology*, 1(3), 262–266. https://doi.org/10.1037/H0021872

Freeman, A. L. J., Kerr, J., Recchia, G., Schneider, C. R., Lawrence, A. C. E., Finikarides, L., Luoni,
G., Dryhurst, S., & Spiegelhalter, D. (2021). Communicating personalized risks from COVID-19:
guidelines from an empirical study. *Royal Society Open Science*, 8(4).
https://doi.org/10.1098/RSOS.201721

Frenda, S. J., Nichols, R. M., & Loftus, E. F. (2011). Current Issues and Advances in Misinformation Research. *Current Directions in Psychological Science*, 20(1), 20–23. https://doi.org/10.1177/0963721410396620

Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). Proceedings of the ... International AAAI Conference on Weblogs and Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, Issue 1). Association for the Advancement of Artificial Intelligence. https://ojs.aaai.org/index.php/ICWSM/article/view/14559

Ganesh, B., & Bright, J. (2020). Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation. *Policy & Internet*, *12*(1), 6–19. https://doi.org/10.1002/POI3.236

Garimella, K., & Eckles, D. (2020). Images and Misinformation in Political Groups: Evidence from WhatsApp in India. https://arxiv.org/pdf/2005.09784.pdf

Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1), 129–149. https://doi.org/10.1111/bjso.12286

Giese, H., Neth, H., Moussaïd, M., Betsch, C., & Gaissmaier, W. (2020). The echo in flu-vaccination echo chambers: Selective attention trumps social influence. *Vaccine*, *38*(8), 2070–2076. https://doi.org/10.1016/J.VACCINE.2019.11.038

Gil de Zúñiga, H., Weeks, B., & Ardèvol-Abreu, A. (2017). Effects of the News-Finds-Me Perception in Communication: Social Media Use Implications for News Seeking and Learning About Politics. *Journal of Computer-Mediated Communication*, 22(3), 105–123. https://doi.org/10.1111/jcc4.12185

Godbold, L. C., & Pfau, M. (2016). Conferring Resistance to Peer Pressure Among Adolescents: Using Inoculation Theory to Discourage Alcohol Use. *Http://Dx.Doi.Org/10.1177/009365000027004001*, 27(4), 411–437. https://doi.org/10.1177/009365000027004001

Goga, O., Loiseau, P., Sommer, R., Teixeira, R., & Gummadi, K. P. (2015). On the Reliability of Profile Matching Across Large Online Social Networks. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1799–1808. https://doi.org/10.1145/2783258.2788601

Goga, O., Venkatadri, G., & Gummadi, K. P. (2015). The doppelgänger bot attack: Exploring identity impersonation in online social networks. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 2015-October, 141–153. https://doi.org/10.1145/2815675.2815699

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

Gottfried, J., & Shearer, E. (2016). *News Use Across Social Media Platforms 2016*. https://www.pewresearch.org/journalism/2016/05/26/news-use-across-social-media-platforms-2016/ Groenendyk, E. (2018). Competing Motives in a Polarized Electorate: Political Responsiveness, Identity Defensiveness, and the Rise of Partisan Antipathy. *Political Psychology*, *39*, 159–171. https://doi.org/10.1111/pops.12481

Groshek, J., & Koc-Michalska, K. (2017). Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign. *Information, Communication & Society*, 20(9), 1389–1407. https://doi.org/10.1080/1369118X.2017.1329334

Gross, B. (2017). Harvesting Social Media for Journalistic Purposes in the UK. In *Privacy, Data Protection and Cybersecurity in Europe* (pp. 31–42). Springer International Publishing. https://doi.org/10.1007/978-3-319-53634-7\_3

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 LP – 15545. https://doi.org/10.1073/pnas.1920498117

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. https://apo.org.au/node/126961

Halpern, D., Valenzuela, S., Katz, J., & Miranda, J. P. (2019). From Belief in Conspiracy Theories to Trust in Others: Which Factors Influence Exposure, Believing and Sharing Fake News. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11578 LNCS, 217–232. https://doi.org/10.1007/978-3-030-21902-4\_16/FIGURES/5

Hansen, N., & Heu, L. (2020). All Human, yet Different: An Emic-Etic Approach to Cross-Cultural Replication in Social Psychology. *Social Psychology*, *51*(6), 361–369. https://doi.org/10.1027/1864-9335/A000436

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias. *Social Cognition*, 27(5), 733–763. https://doi.org/10.1521/soco.2009.27.5.733

Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, 6(1), 38. https://doi.org/10.1186/s41235-021-00301-5

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature 2010* 466:7302, 466(7302), 29–29. https://doi.org/10.1038/466029a

Hinsley, A., & Holton, A. (2021). Fake News Cues: Examining the Impact of Content, Source, and Typology of News Cues on People's Confidence in Identifying Mis- and Disinformation . *International Journal of Communication*, *15*. https://ijoc.org/index.php/ijoc/article/view/12387

Hinsley, A., Ju, I., Park, T., & Ohs, J. (2022). Credibility in the time of COVID-19: Cues that audiences look for when assessing information on social media and building confidence in identifying 'fake news' about the virus. *Open Information Science*, *6*(1), 61–73. https://doi.org/10.1515/OPIS-2022-0132

Holland, R. W., Verplanken, B., & van Knippenberg, A. (2003). From repetition to conviction: Attitude accessibility as a determinant of attitude certainty. *Journal of Experimental Social Psychology*, *39*(6), 594–601. https://doi.org/10.1016/S0022-1031(03)00038-6

Holyoak, K. J., & Thagard, P. (1995). Mental Leaps: Analogy in Creative Thought.

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association, 37*(4), 307–315. https://doi.org/10.1037/hea0000586

Hoseini, M., Melo, P., Benevenuto, F., Feldmann, A., & Zannettou, S. (2021). *On the Globalization of the QAnon Conspiracy Theory Through Telegram*. https://doi.org/10.48550/arxiv.2105.13020

Hunter, J., Levine, R., & Sayers, S. (1976). Attitude Change in Hierarchical Belief Systems and Its Relationship to Persuasibility, Dogmatism, and Rigidity. *Human Communication Research*, *3*(1), 3–28. https://doi.org/10.1111/J.1468-2958.1976.TB00501.X

Imhoff, R., & Lamberty, P. (2020). A bioweapon or a hoax? The link between distinct conspiracy beliefs about the Coronavirus disease (COVID-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, *11*(8), 1110–1118. https://doi.org/10.1177/1948550620934692

Ivanov, B., Miller, C. H., Compton, J., Averbeck, J. M., Harrison, K. J., Sims, J., Parker, K. A., & Parker, J. L. (2012). Effects of Postinoculation Talk on Resistance to Influence. *Journal of Communication*, 62(4), 701–718.

Ivanov, B., Parker, K. A., & Dillingham, L. L. (2018). Testing the Limits of Inoculation-Generated Resistance. *Western Journal of Communication*, 82(5), 648–665. https://doi.org/10.1080/10570314.2018.1454600

Ivanov, B., Pfau, M., & Parker, K. A. (2009). The Attitude Base as a Moderator of the Effectiveness of Inoculation Strategy. *Communication Monographs*, 76(1), 47–72. https://doi.org/10.1080/03637750802682471

Ivanov, B., Rains, S. A., Geegan, S. A., Vos, S. C., Haarstad, N. D., & Parker, K. A. (2017). Beyond Simple Inoculation: Examining the Persuasive Value of Inoculation for Audiences with Initially Neutral or Opposing Attitudes. *Western Journal of Communication*, *81*(1), 105–126. https://doi.org/10.1080/10570314.2016.1224917

Ivanov, B., Sellnow, T., Getchell, M., & Burns, W. (2018). The potential for inoculation messages and postinoculation talk to minimize the social impact of politically motivated acts of violence. *Journal of Contingencies and Crisis Management*, *26*(4), 414–424. https://doi.org/10.1111/1468-5973.12213

Iyengar, S., & Krupenkin, M. (2018). The Strengthening of Partisan Affect. *Political Psychology*, *39*, 201–218. https://doi.org/10.1111/pops.12487

Iyengar, S., & Massey, D. S. (2019). Scientific communication in a post-truth society. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(16), 7656–7661. https://doi.org/10.1073/PNAS.1805868115

Jalil, A., Tohara, T., Shuhidan, S. M., Diana, F., Bahry, S., & Norazmi Bin Nordin, M. (2021). Exploring Digital Literacy Strategies for Students with Special Educational Needs in the Digital Age. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(9), 3345–3358. https://doi.org/10.17762/TURCOMAT.V12I9.5741

Jessen, J., & Jørgensen, A. H. (2012). Aggregated trustworthiness: Redefining online credibility through social validation. *First Monday*, *17*(1). https://doi.org/10.5210/fm.v17i1.3731

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

Jolley, D., & Paterson, J. L. (2020a). Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology*. https://doi.org/10.1111/bjso.12394

Jolley, D., & Paterson, J. L. (2020b). Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology*, *59*(3), 628–640. https://doi.org/10.1111/bjso.12394

Jung, A. M. (2011). Twittering Away the Right of Publicity: Personality Rights and Celebrity Impersonation on Social Networking Websites. *Chicago-Kent Law Review*, 86. https://heinonline.org/HOL/Page?handle=hein.journals/chknt86&id=389&div=18&collection=journ als Kabha, R., Kamel, A., & Elbahi, M. (2019). Comparison Study between the UAE, the UK, and India in Dealing with WhatsApp Fake News. *Community & Communication Amity School of Communication*, *10*, 2456–9011. https://doi.org/10.31620/JCCC.12.19/18

Kalra, A., & Vengattil, M. (2019). WhatsApp threatens legal action against public claims of messaging abuses / Reuters. https://www.reuters.com/article/us-whatsapp-abuse-idUSKCN1TC1V1

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kazemi, A., Garimella, K., Shahi, G. K., Gaffney, D., & Hale, S. A. (2021). Tiplines to Combat Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian Election on WhatsApp. http://arxiv.org/abs/2106.04726

Konijn, E. (2013). The role of emotion in media use and effects. In *The Oxford handbook of media psychology* (pp. 186–211). Oxford University Press. https://psycnet.apa.org/record/2013-00995-011 Kucharski, A. (2016). Study epidemiology of fake news. *Nature*, *540*(7634), 525–525.

https://doi.org/10.1038/540525a

Kucharski, A. (2020, February 8). Misinformation on the coronavirus might be the most contagious thing about it. *The Guardian*.

Lanius, C., Weber, R., & MacKenzie, W. I. (2021). Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey. *Social Network Analysis and Mining*, *11*(1), 1–15. https://doi.org/10.1007/S13278-021-00739-X/FIGURES/7

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. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lazić, A., & Žeželj, I. (2021). A systematic review of narrative interventions: Lessons for countering anti-vaccination conspiracy theories and misinformation. *Public Understanding of Science*, *30*(6), 644–670. https://doi.org/10.1177/09636625211011881

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/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., & van der Linden, S. (2021a). Countering Misinformation and Fake News Through Inoculation and Prebunking. *Https://Doi.Org/10.1080/10463283.2021.1876983*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lewandowsky, S., & van der Linden, S. (2021b). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 1–38. 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(1), 1–15. https://doi.org/10.1186/S41235-021-00323-Z/FIGURES/4

Lim, J. S., & Ki, E.-J. (2007). Resistance to Ethically Suspicious Parody Video on YouTube: A Test of Inoculation Theory. *Journalism & Mass Communication Quarterly*, 84(4), 713–728. https://doi.org/10.1177/107769900708400404

Lin, W.-K., & Pfau, M. (2007). Can Inoculation Work Against the Spiral of Silence? A Study of Public Opinion on the Future of Taiwan. *International Journal of Public Opinion Research*, *19*(2), 155–172. https://doi.org/10.1093/ijpor/edl030

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 2021 5:3*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

Lumsdaine, A. A., & Janis, I. L. (1953). Resistance to "Counterpropaganda" Produced by One-Sided and Two-Sided "Propaganda" Presentations. *Public Opinion Quarterly*, *17*(3), 311–318. https://doi.org/10.1086/266464

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(23). https://doi.org/10.1073/PNAS.2019527118

Lyons, M. T., & Hughes, S. (2015). Malicious mouths? The Dark Triad and motivations for gossip. *Personality and Individual Differences*, 78, 1–4. https://doi.org/10.1016/J.PAID.2015.01.009

MacFarlane, D., Tay, L. Q., Hurlstone, M. J., & Ecker, U. (2020). *Refuting Spurious COVID-19 Treatment Claims Reduces Demand and Misinformation Sharing*. https://doi.org/10.31234/OSF.IO/Q3MKD

Machado, C., Kira, B., Narayanan, V., Kollanyi, B., & Howard, P. N. (2019). A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections. *Companion Proceedings of The 2019 World Wide Web Conference*. https://doi.org/10.1145/3308560

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

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2020). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*. https://doi.org/https://dx.doi.org/10.1037/xap0000315

Mandical, R. R., Mamatha, N., Shivakumar, N., Monica, R., & Krishna, A. N. (2020). Identification of Fake News Using Machine Learning. *Proceedings of CONECCT 2020 - 6th IEEE International Conference on Electronics, Computing and Communication Technologies*. https://doi.org/10.1109/CONECCT50063.2020.9198610

Mauet, T. (1992). *Fundamentals of Trial Techniques* . Little Brown. https://www.ojp.gov/ncjrs/virtual-library/abstracts/fundamentals-trial-techniques

McCroskey, J. C., Young, T. J., & Scott, M. D. (1972). The effects of message sidedness and evidence on inoculation against counterpersuasion in small group communication. *Http://Dx.Doi.Org/10.1080/03637757209375758*, *39*(3), 205–212. https://doi.org/10.1080/03637757209375758

McElhaney, J. (1987). *McElhaney's trial notebook*. Section of Litigation, American Bar Association. https://books.google.com/books/about/McElhaney\_s\_Trial\_Notebook.html?hl=de&id=-

fH4AAAAIAAJ

McGuire, W. J. (1964a). Inducing resistance against persuasion: Some Contemporary Approaches. *Advances* in *Experimental Social Psychology*, *1*, 191–229. https://doi.org/http://dx.doi.org/10.1016/S0065-2601(08)60052-0

McGuire, W. J. (1964b). *Some Contemporary Approaches* (pp. 191–229). https://doi.org/10.1016/S0065-2601(08)60052-0

McGuire, W. J., & Papageorgis, D. (1961a). 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.

McGuire, W. J., & Papageorgis, D. (1961b). The relative efficacy of various types of prior beliefdefense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology*, *62*(2), 327–337. https://doi.org/10.1037/h0042026

Mcintryre, lee. (2018). Post-Thruth. MIT Press. https://mitpress.mit.edu/books/post-truth

Mena, P., Barbe, D., & Chan-Olmsted, S. (2020). Misinformation on Instagram: The Impact of Trusted Endorsements on Message Credibility. *Social Media* + *Society*, *6*(2), 205630512093510. https://doi.org/10.1177/2056305120935102

Miller, C. H., Ivanov, B., Sims, J., Compton, J., Harrison, K. J., Parker, K. A., Parker, J. L., & Averbeck, J. M. (2013). Boosting the Potency of Resistance: Combining the Motivational Forces of Inoculation and Psychological Reactance. *Human Communication Research*, *39*(1), 127–155. https://doi.org/10.1111/j.1468-2958.2012.01438.x

Miller, G. R., & Burgoon, M. (1972). *New techniques of persuasion*. Harper & Row. https://books.google.com/books/about/New\_Techniques\_of\_Persuasion.html?hl=de&id=\_zwlAQAA IAAJ

Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(18), 5631–5636. https://doi.org/10.1073/pnas.1421883112

Mukherjee, R. (2020). Mobile witnessing on WhatsApp: Vigilante virality and the anatomy of mob lynching. *Https://Doi.Org/10.1080/14746689.2020.1736810*, *18*(1), 79–101. https://doi.org/10.1080/14746689.2020.1736810

Murphy, J., Keane, A., & Power, A. (2020). Computational propaganda: Targeted advertising and the perception of truth. *European Conference on Information Warfare and Security, ECCWS, 2020-June,* 491–500. https://doi.org/10.34190/EWS.20.503

Nabi, R. L. (2003a). "Feeling" Resistance: Exploring the Role of Emotionally Evocative Visuals in Inducing Inoculation. *Media Psychology*, 5(2), 199–223. https://doi.org/10.1207/S1532785XMEP0502\_4

Nabi, R. L. (2003b). "Feeling Resistance": Exploring the Role of Emotionally Evocative Visuals in Inducing Inoculation. *Media Psychology*, 5(2), 199–223. https://doi.org/10.1207/S1532785XMEP0502\_4

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report*. SSRN Electronic Journal. https://ssrn.com/abstract=3414941

Newman, N., Fletcher, R., Schulz, A., Andi, S., & Nielsen, R. K. (2020). Reuters Institute DigitalNewsReport2020.Reutersinstitute.Politics.Ox.Ac.Uk.

https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR\_2020\_FINAL.pdf

Nisbet, E. C., Cooper, K. E., & Garrett, R. K. (2015). The Partisan Brain: How Dissonant Science Messages Lead Conservatives and Liberals to (Dis)Trust Science. *Https://Doi.Org/10.1177/0002716214555474*, 658(1), 36–66. https://doi.org/10.1177/0002716214555474

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

O'Keefe, D. J. (2015). Message generalizations that support evidence-based persuasive message design: Specifying the evidentiary requirements. *Health Communication*, *30*(2), 106–113.

Oliva, T. D. (2020). Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. *Human Rights Law Review*, 20(4), 607–640. https://doi.org/10.1093/HRLR/NGAA032

Papageorgis, D., & McGuire, W. J. (1961). The generality of immunity to persuasion produced by pre-exposure to weakened counterarguments. *The Journal of Abnormal and Social Psychology*, 62(3), 475–481. https://doi.org/10.1037/h0048430

Park, C. S. (2019). Does Too Much News on Social Media Discourage News Seeking? Mediating Role of News Efficacy Between Perceived News Overload and News Avoidance on Social Media. *Social Media* + *Society*, *5*(3), 205630511987295. https://doi.org/10.1177/2056305119872956

Parker, K. A., Ivanov, B., & Compton, J. (2012a). Inoculation's Efficacy With Young Adults' Risky Behaviors: Can Inoculation Confer Cross-Protection Over Related but Untreated Issues? *Health Communication*, 27(3), 223–233. https://doi.org/10.1080/10410236.2011.575541

Parker, K. A., Ivanov, B., & Compton, J. (2012b). Inoculation's Efficacy With Young Adults' Risky Behaviors: Can Inoculation Confer Cross-Protection Over Related but Untreated Issues? *Health Communication*, 27(3), 223–233. https://doi.org/10.1080/10410236.2011.575541

Parker, K. A., Rains, S. A., & Ivanov, B. (2016). Examining the "Blanket of Protection" Conferred by Inoculation: The Effects of Inoculation Messages on the Cross-protection of Related Attitudes. *Communication Monographs*, *83*(1), 49–68. https://doi.org/10.1080/03637751.2015.1030681

Parrott, R., Silk, K., Dorgan, K., Condit, C., & Harris, T. (2005). Risk Comprehension and Judgments of Statistical Evidentiary Appeals. *Human Communication Research*, *31*(3), 423–452. https://doi.org/10.1111/J.1468-2958.2005.TB00878.X

Peachham, H. (1577). *The Garden Of Eloquence: Conteyning the Figures Of Grammer and Rhetorick*. https://www.proquest.com/publication/2066892?OpenUrlRefId=info:xri/sid:primo&parentSessionId =Mv9ZlsWG5JRTH%2F3tkDBAMf56CtHZP2ejG%2F37YYAatpU%3D&accountid=9851

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017a). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017b). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. https://doi.org/10.1016/J.JESP.2017.01.006

Pennycook, G., Cannon, T., & Rand, D. G. (2018). Prior Exposure Increases Perceived Accuracy of Fake News. *Journal of Experimental Psychology*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 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. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88(2), 185–200. https://doi.org/10.1111/jopy.12476

Perel, M., & Elkin-Koren, N. (2015). Accountability in Algorithmic Copyright Enforcement. StanfordTechnologyLawReview,19.https://heinonline.org/HOL/Page?handle=hein.journals/stantlr19&id=489&div=20&collection=journals

Peter, C., & Koch, T. (2015). When Debunking Scientific Myths Fails (and When It Does Not): The Backfire Effect in the Context of Journalistic Coverage and Immediate Judgments as Prevention Strategy. *Science Communication*, *38*(1), 3–25. https://doi.org/10.1177/1075547015613523

Peters, U. (2020a). What Is the Function of Confirmation Bias? *Erkenntnis* 2020 87:3, 87(3), 1351–1376. https://doi.org/10.1007/S10670-020-00252-1

Peters, U. (2020b). Illegitimate Values, Confirmation Bias, and Mandevillian Cognition in Science. *Https://Doi.Org/10.1093/Bjps/Axy079*, 72(4), 1061–1081. https://doi.org/10.1093/BJPS/AXY079 Petersen, A. M., Vincent, E. M., & Westerling, A. L. (2019). Discrepancy in scientific authority and media visibility of climate change scientists and contrarians. *Nature Communications*, *10*(1), 3502. https://doi.org/10.1038/s41467-019-09959-4

Petrocelli, J. v, & Rucker, D. D. (2007). Unpacking Attitude Certainty: Attitude Clarity and Attitude Correctness Bullshitting View project Ease of Retrieval Moderates the Effects of Power: Implications for the Replicability of Power Recall Effects View project. *Article in Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.92.1.30

Petrocelli, J. v., Tormala, Z. L., & Rucker, D. D. (2007). Unpacking attitude certainty: Attitude clarity and attitude correctness. *Journal of Personality and Social Psychology*, 92(1), 30–41. https://doi.org/10.1037/0022-3514.92.1.30

Petty, R. E., & Krosnick, J. A. (1995). *Attitude strength: antecedents and consequences*. https://books.google.co.uk/books?hl=en&lr=&id=taWYAgAAQBAJ&oi=fnd&pg=PA413&dq=Eagl y,+A.,+%26+Chaiken,+S.+(1995).+The+psychology+of+attitudes&ots=wAFstpldxZ&sig=QDjzhjy dgkhgYAcotsi6jQVcnQI#v=onepage&q=Eagly%2C%20A.%2C%20%26%20Chaiken%2C%20S.% 20(1995).%20The%20psychology%20of%20attitudes&f=false

Petty, R. E., & Wegener, D. T. (1998). Attitude Chnnge: Multiple Roles for Persuasion Variables. *The Handbook of Social Psychology*.

Pfau, M., Bockern, S. van, & Kang, J. G. (1992). Use of inoculation to promote resistance to smoking initiation among adolescents. *Communication Monographs*, *59*(3), 213–230. https://doi.org/10.1080/03637759209376266

Pfau, M., & Burgoon, M. (1988). Inoculation in Political Campaign Communication. *Human Communication Research*, *15*(1), 91–111. https://doi.org/10.1111/J.1468-2958.1988.TB00172.X

Pfau, M., Compton, J., Parker, K. A., An, C., Wittenberg, E. M., Ferguson, M., Horton, H., & Malyshev, Y. (2006). The Conundrum of the Timing of Counterarguing Effects in Resistance: Strategies to Boost the Persistence of Counterarguing Output. *Communication Quarterly*, *54*(2), 143–156. https://doi.org/10.1080/01463370600650845

Pfau, M., Holbert, R. L., Zubric, S. J., Pasha, N. H., & Lin, W.-K. (2000). Role and Influence of Communication Modality in the Process of Resistance to Persuasion. *Media Psychology*, 2(1), 1–33. https://doi.org/10.1207/S1532785XMEP0201\_1

Pfau, M., Ivanov, B., Houston, B., Haigh, M., Sims, J., Gilchrist, E., Russell, J., Wigley, S., Eckstein, J., & Richert, N. (2005). Inoculation and Mental Processing: The Instrumental Role of Associative Networks in the Process of Resistance to Counterattitudinal Influence. *Communication Monographs*, 72(4), 414–441. https://doi.org/10.1080/03637750500322578

Pfau, M., Szabo, A., Anderson, J., Morrill, J., Zubric, J., & H-Wan, H.-H. (2001). The role and impact of affect in the process of resistance to persuasion. *Human Communication Research*, 27(2), 216–252. https://doi.org/10.1111/j.1468-2958.2001.tb00781.x

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A Motivational Model of Video Game Engagement. *Review of General Psychology*, *14*(2), 154–166. https://doi.org/10.1037/a0019440

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/https://doi.org/10.1016/j.dr.2016.06.004

Rasi, P., Vuojärvi, H., & Ruokamo, H. (2019). Media Literacy Education for All Ages. *Journal of Media Literacy Education*, *11*(2), 1–19. https://doi.org/https://doi.org/10.23860/JMLE-2019-11-2-1

Rathje, S., van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences of the United States of America*, 118(26).

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J. M., Benevenuto, F., & Vas-Concelos, M. (2019). (*Mis*)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. 11.

Reznik, M. (2012). Identity Theft on Social Networking Sites: Developing Issues of Internet Impersonation. *Touro Law Review*, 29.

Rhodes, S. C. (2022). Filter Bubbles, Echo Chambers, and Fake News: How Social Media Conditions Individuals to Be Less Critical of Political Misinformation. *Political Communication*, *39*(1), 1–22. https://doi.org/10.1080/10584609.2021.1910887

Richards, A. S., & Banas, J. A. (2018a). The Opposing Mediational Effects of Apprehensive Threat and Motivational Threat When Inoculating Against Reactance to Health Promotion. *Southern Communication Journal*, 83(4), 245–255. https://doi.org/10.1080/1041794X.2018.1498909

Richards, A. S., & Banas, J. A. (2018b). The Opposing Mediational Effects of Apprehensive Threat and Motivational Threat When Inoculating Against Reactance to Health Promotion. *Southern Communication Journal*, *83*(4), 245–255. https://doi.org/10.1080/1041794X.2018.1498909

Rodny-Gumede, Y. (2018). Fake It till You Make It: The Role, Impact and Consequences of Fake News. In *Perspectives on Political Communication in Africa* (pp. 203–219). Springer International Publishing. https://doi.org/10.1007/978-3-319-62057-2\_13

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2020). Differentiating Item and Testing Effects in Inoculation Research on Online Misinformation. *Educational and Psychological Measurement*, 1–23. https://doi.org/10.1177/0013164420940378

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(2011199). https://doi.org/10.1098/rsos.201199

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. (2019a). 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. (2019b). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1), 65. https://doi.org/10.1057/s41599-019-0279-9

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

Rosnow, R. L., Esposito, J. L., & Gibney, L. (1988). Factors influencing rumor spreading: Replication and extension. *Language and Communication*, 8(1), 29–42. https://doi.org/10.1016/0271-5309(88)90004-3

Rucker, D. D., & Petty, R. E. (2004a). When Resistance Is Futile: Consequences of Failed Counterarguing for Attitude Certainty. *Journal of Personality and Social Psychology*, 86(2), 219–235. https://doi.org/10.1037/0022-3514.86.2.219

Rucker, D. D., & Petty, R. E. (2004b). When resistance is futile: consequences of failed counterarguing for attitude certainty. *Journal of Personality and Social Psychology*, 86(2), 219–235. https://doi.org/10.1037/0022-3514.86.2.219

Runciman, D. (2016). *Is this how democracy ends?* 38(23). https://www.lrb.co.uk/the-paper/v38/n23/david-runciman/is-this-how-democracy-ends

Saleh, N., Roozenbeek, J., Makki, F., McClanahan, W., & van der Linden, S. (2020). Active inoculation boosts attitudinal resistance against extremist persuasion techniques – A novel approach towards the prevention of violent extremism. *Behavioural Public Policy*. https://doi.org/10.1017/bpp.2020.60

Sandel, M. J. (1998). *Democracy's discontent : America in search of a public philosophy*. Belknap Press of Harvard University Press.

Scales, D., Gorman, J., & Jamieson, K. H. (2021). The Covid-19 Infodemic — Applying the Epidemiologic Model to Counter Misinformation. *New England Journal of Medicine*, *385*(8), 678–681.

Schaeffer, K. (2020, April 8). Nearly three-in-ten Americans believe COVID-19 was made in a lab. *Pew Research Center*.

Schiefer, D., & van der Noll, J. (2017). The Essentials of Social Cohesion: A Literature Review. *Social Indicators Research*, *132*(2), 579–603. https://doi.org/10.1007/S11205-016-1314-5/FIGURES/2

Seo, Y., Kim, J., Choi, Y. K., & Li, X. (2019). In "likes" we trust: likes, disclosures and firm-serving motives on social media. *European Journal of Marketing*, 53(10), 2173–2192. https://doi.org/10.1108/EJM-11-2017-0883

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, 22, 100104. https://doi.org/10.1016/J.OSNEM.2020.100104

Simon, F. M., & Camargo, C. Q. (2021). Autopsy of a metaphor: The origins, use and blind spots of the 'infodemic': *Https://Doi.Org/10.1177/14614448211031908*. https://doi.org/10.1177/14614448211031908

Soroya, S. H., Farooq, A., Mahmood, K., Isoaho, J., & Zara, S. (2021). From information seeking to information avoidance: Understanding the health information behavior during a global health crisis. *Information Processing & Management*, *58*(2), 102440. https://doi.org/10.1016/J.IPM.2020.102440 Southwell, B. G., & Yzer, M. C. (2009). When (and Why) Interpersonal Talk Matters for Campaigns. *Communication Theory*, *19*(1), 1–8. https://doi.org/10.1111/j.1468-2885.2008.01329.x

 Sparks, C. (2014). What is the "Digital Divide" and why is it Important?

 *Https://Doi.Org/10.1080/13183222.2013.11009113*,
 20(2),
 27–46.

 https://doi.org/10.1080/13183222.2013.11009113
 20(2),
 27–46.

Steenbuch Traberg, C. (2022). Misinformation: broaden definition to curb its societal influence. *Nature*, 606(7915), 653–653. https://doi.org/10.1038/D41586-022-01700-4

Sun, T., Youn, S., Wu, G., & Kuntaraporn, M. (2006). Online Word-of-Mouth (or Mouse): An Exploration of Its Antecedents and Consequences. *Journal of Computer-Mediated Communication*, *11*(4), 1104–1127. https://doi.org/10.1111/j.1083-6101.2006.00310.x

Szpitalak, M., & Polczyk, R. (2014). Mental fatigue, mental warm-up, and self-reference as determinants of the misinformation effect., *25*(2), 135–151.

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, *51*, 72–82. https://doi.org/10.1016/J.JRETCONSER.2019.05.026

Tormala, Z. (2015). The role of certainty (and uncertainty) in attitudes and persuasion. https://doi.org/10.1016/j.copsyc.2015.10.017

Tormala, Z. L. (2016). The role of certainty (and uncertainty) in attitudes and persuasion. *Current Opinion in Psychology*, *10*, 6–11. https://doi.org/https://doi.org/10.1016/j.copsyc.2015.10.017

Tormala, Z. L., Clarkson, J. J., & Petty, R. E. (2006). Resisting persuasion by the skin of one's teeth: The hidden success of resisted persuasive messages. *Journal of Personality and Social Psychology*, *91*(3), 423–435. https://doi.org/10.1037/0022-3514.91.3.423

Tormala, Z. L., DeSensi, V. L., & Petty, R. E. (2007). Resisting Persuasion by Illegitimate Means: A Metacognitive Perspective on Minority Influence. *Personality and Social Psychology Bulletin*, *33*(3), 354–367. https://doi.org/10.1177/0146167206295004

Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me makes me stronger: the effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, *83*(6), 1298–1313. https://doi.org/10.1037//0022-3514.83.6.1298

Tormala, Z. L., & Rucker, D. D. (2018). Attitude certainty: Antecedents, consequences, and new directions. *Consumer Psychology Review*, *1*(1), 72–89. https://doi.org/10.1002/ARCP.1004

Tormala, Z., & Petty, R. (2002). What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, *83*(6), 1298–1313. https://doi.org/10.1037/0022-3514.83.6.1298

Tormala, Z., & Petty, R. (2004). Source Credibility and Attitude Certainty: A Metacognitive Analysis of Resistance to Persuasion. *Journal of Consumer Psychology*, *14*(4), 427–442. https://doi.org/10.1207/s15327663jcp1404\_11 Tormala, Z., & Rucker, D. (2015). *How Certainty Transforms Persuasion*. Harvard Business Review. https://hbr.org/2015/09/how-certainty-transforms-persuasion

Touré-Tillery, M., & Mcgill, A. L. (2015). Who or What to Believe: Trust and the Differential Persuasiveness of Human and Anthropomorphized Messengers. *Journal of Marketing*, *79*, 1547–7185. https://journals.sagepub.com/doi/pdf/10.1509/jm.12.0166

Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, *185*, 111269. https://doi.org/10.1016/J.PAID.2021.111269

UNESCO. (2020). *#ThinkBeforeSharing - Stop the spread of conspiracy theories*. En.Unesco.Org. https://en.unesco.org/themes/gced/thinkbeforesharing

Urman, A., & Katz, S. (2020). What they do in the shadows: examining the far-right networks on Telegram. *Information Communication and Society*. https://doi.org/10.1080/1369118X.2020.1803946/SUPPL\_FILE/RICS\_A\_1803946\_SM9952.GEXF U.S. Department of and Education. (2017). *Highlights of U.S. National Results*. https://nces.ed.gov/surveys/piaac/national\_results.asp

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., ... Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z

van der Linden, S. (2017). The nature of viral altruism and how to make it stick. *Nature Human Behaviour 2017 1:3, 1*(3), 1–4. https://doi.org/10.1038/s41562-016-0041

van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine 2022 28:3*, *28*(3), 460–467. 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. (2020). Psychological inoculation against fake news. In R. Greifenader, M. Jaffé, E. Newman, & N. Schwarz (Eds.), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*. Psychology Press. https://doi.org/10.4324/9780429295379-11

van der Linden, S., Roozenbeek, J., & Compton, J. A. (2020). Inoculating Against Fake News About COVID-19. *Frontiers in Psychologyi*, *11*(566790). https://doi.org/10.3389/fpsyg.2020.566790

Visser, P. S., Krosnick, J. A., & Simmons, J. P. (2003). Distinguishing the cognitive and behavioral consequences of attitude importance and certainty: A new approach to testing the common-factor hypothesis. *Journal of Experimental Social Psychology*, *39*(2), 118–141. https://doi.org/10.1016/S0022-1031(02)00522-X

Vivion, M., Sidi, E. A. L., Betsch, C., Dionne, M., Dubé, E., Driedger, S. M., Gagnon, D., Graham, J., Greyson, D., Hamel, D., Lewandowsky, S., MacDonald, N., Malo, B., Meyer, S. B., Schmid, P.,

Steenbeek, A., Linden, S. van der, Verger, P., Witteman, H. O., ... (CIRN), C. I. R. N. (2022). Prebunking messaging to inoculate against COVID-19 vaccine misinformation: an effective strategy for public health. *Https://Doi.Org/10.1080/17538068.2022.2044606*, 1–11. https://doi.org/10.1080/17538068.2022.2044606

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. (2020a). 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. (2020b). 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

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

Williams, B. A., & Delli Carpini, M. X. (2016). Monica and Bill All the Time and Everywhere: The Collapse of Gatekeeping and Agenda Setting in the New Media Environment. *Http://Dx.Doi.Org/10.1177/0002764203262344*, 47(9), 1208–1230. https://doi.org/10.1177/0002764203262344

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

Wolf, F., Lorenz-Spreen, P., & Lehmann, S. (2021). Successive cohorts of Twitter users show increasing activity and shrinking content horizons. https://arxiv.org/pdf/2108.08641.pdf

Wood. (2007). Rethinking the Inoculation Analogy: Effects on Subjects With Differing Preexisting Attitudes. *Human Communication Research*, *33*(3), 357–378. https://doi.org/10.1111/j.1468-2958.2007.00303.x

Wood, M. (2007a). Rethinking the Inoculation Analogy: Effects on Subjects With Differing Preexisting Attitudes. *Human Communication Research*, *33*(3), 357–378. https://doi.org/10.1111/j.1468-2958.2007.00303.x

Wood, M. (2007b). Rethinking the Inoculation Analogy: Effects on Subjects With Differing Preexisting Attitudes. *Human Communication Research*, *33*(3), 357–378. https://doi.org/10.1111/J.1468-2958.2007.00303.X

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

World Health Organization. (2020). *Coronavirus disease (COVID-19) advice for the public: Mythbusters*. Www.Who.Int. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters Wyer, R. S. (1976). Effects of previously formed beliefs on syllogistic inference processes. *Journal of Personality and Social Psychology*, *33*(3), 307–316. https://doi.org/10.1037/0022-3514.33.3.07

Yasmin, S. (2021). *Viral BS: medical myths and why we fall for them*. https://www.amazon.com/Viral-BS-Medical-Myths-Fall/dp/1421440407

Zarocostas, J. (2020a). How to fight an infodemic. *The Lancet*, 395(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Zarocostas, J. (2020b). How to fight an infodemic. *Lancet (London, England)*, 395(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Zerback, T., Töpfl, F., & Knöpfle, M. (2020). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them: *Https://Doi.Org/10.1177/1461444820908530*, 23(5), 1080–1098. https://doi.org/10.1177/1461444820908530

Zhao, Y., & Zhang, J. (2017). Consumer health information seeking in social media: a literature review. *Health Information & Libraries Journal*, *34*(4), 268–283. https://doi.org/10.1111/hir.12192 Zhou, Y., & Shen, L. (2021). Confirmation Bias and the Persistence of Misinformation on Climate Change. *Communication Research*, 009365022110280. https://doi.org/10.1177/00936502211028049 Zhu, B., Chen, C., Loftus, E. F., Lin, C., He, Q., Chen, C., Li, H., Moyzis, R. K., Lessard, J., & Dong, Q. (2010). Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. *Personality and Individual Differences*, *48*(8), 889–894. https://doi.org/10.1016/J.PAID.2010.02.016

Zipursky, R. (2019). Nuts About NETZ: The Network Enforcement Act and Freedom of Expression. In *Fordham International Law Journal* (Vol. 42, Issue 4). https://ir.lawnet.fordham.edu/ilj

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), e0138740. https://doi.org/10.1371/journal.pone.0138740

Zuwerink, J. R., & Devine, P. G. (1996). Attitude Importance and Resistance to Persuasion: It's Not Just the Thought That Counts. *Journal of Personality and Social Psychology*, *70*(5), 931–944. https://doi.org/10.1037/0022-3514.70.5.9s31

## **8** APPENDICES

Appendix 1: Average reliability (pre-post) judgments overall and for each fake news badge by experimental condition. 199 Appendix 2: Average confidence (pre-post) judgments overall and for each fake news badge by experimental condition. 199 Appendix 4: All 18 chat screenshots (fake, real news items) participants viewed pre-post by badge. 201 Appendix 5: Appendix 1: Average reliability (pre-post) judgments for each fake news badge by experimental condition. 201 Appendix 6:Linear regression for demographics. 202 Appendix 7: Attack message on fictitious chemical presented at T1. 203 Appendix 8: Inoculation treatment used at T1. 204 Appendix 9: Control message on fictitious chemical used at T1. 204 Appendix 11: Visualisation of the 'Degrees of Resistance' Index. 205 Appendix 10: Intercoder reliability analyses for content measures. 205 Appendix 12: PIT messages (T2) as the vicarious inoculation treatment. 206 Appendix 13: PIT messages (T2) that scored lowest on Resistance Index and were used instead of a traditional attack message. 208 Appendix 14: Fictitious, neutral, and unrelated treatment for the control condition. 208

Chapter 2

Appendix 1: Average reliability (pre-post) judgments overall and for each fake news badge by experimental condition.

**Experimental condition** 

|                 | <b>Inoculation</b> (n = 96) |                |                | Co                    | ontrol (n =   | = 102)            |               |                       |           |
|-----------------|-----------------------------|----------------|----------------|-----------------------|---------------|-------------------|---------------|-----------------------|-----------|
|                 | M <sub>pre</sub>            | $M_{\rm post}$ | $M_{\rm diff}$ | 95%CI <sub>diff</sub> | $M_{\rm pre}$ | M <sub>post</sub> | $M_{ m diff}$ | 95%CI <sub>diff</sub> | Cohen's d |
| Fake news scale | 3.14                        | 2.69           | -0.45          | [-0.29, -0.61]        | 3.32          | 3.23              | -0.09         | [-0.03, -0.15]        | 0.60      |
| Impersonation   | 3.22                        | 2.76           | -0.46          | [-0.21, -0.70]        | 3.49          | 3.48              | -0.01         | [-0.16, 0.14]         | 0.45      |
| Polarisation    | 2.85                        | 2.59           | -0.26          | [-0.03, -0.48]        | 3.07          | 2.95              | -0.12         | [-0.27, 0.02]         | 0.14      |
| Conspiracy      | 3.13                        | 2.58           | -0.55          | [-0.33, -0.77]        | 3.47          | 3.27              | -0.20         | [-0.04, -0.36]        | 0.36      |
| Emotion         | 3.39                        | 2.87           | -0.52          | [-0.29, -0.74]        | 3.53          | 3.44              | -0.09         | [-0.23, 0.05]         | 0.45      |
| Discrediting    | 3.36                        | 2.80           | -0.56          | [-0.33, -0.79]        | 3.39          | 3.41              | 0.02          | [-0.19, 0.14]         | 0.58      |
| Trolling        | 2.83                        | 2.52           | -0.31          | [-0.12, -0.49]        | 2.96          | 2.83              | -0.13         | [-0.02, 0.28]         | 0.22      |

Appendix 2: Average confidence (pre-post) judgments overall and for each fake news badge by experimental condition.

| Experimental condition |                             |                   |                |                       |               |                   |                |                        |           |
|------------------------|-----------------------------|-------------------|----------------|-----------------------|---------------|-------------------|----------------|------------------------|-----------|
|                        | <b>Inoculation</b> (n = 96) |                   |                |                       | Con           | trol (n=          | 102)           |                        |           |
|                        | M <sub>pre</sub>            | M <sub>post</sub> | $M_{\rm diff}$ | 95%CI <sub>diff</sub> | $M_{\rm pre}$ | M <sub>post</sub> | $M_{\rm diff}$ | 95% CI <sub>diff</sub> | Cohen's d |
| Fake news scale        | 5.25                        | 5.47              | 0.22           | [0.10, 0.34]          | 5.27          | 5.21              | -0.06          | [-0.03, 0.14]          | 0.52      |
| Impersonation          | 5.56                        | 5.71              | 0.15           | [-0.05, 0.35]         | 5.68          | 5.51              | -0.17          | [-0.00, 0.34]          | 0.34      |
| Polarisation           | 5.13                        | 5.34              | 0.21           | [0.01, 0.41]          | 5.11          | 5.10              | -0.01          | [-0.14, 0.16]          | 0.25      |
| Conspiracy             | 5.14                        | 5.40              | 0.26           | [0.08, 0.44]          | 5.10          | 5.12              | 0.02           | [-0.20, 0.15]          | 0.27      |
| Emotion                | 5.15                        | 5.41              | 0.26           | [0.10, 0.42]          | 5.25          | 5.12              | -0.13          | [-0.03, 0.29]          | 0.49      |
| Discrediting           | 5.28                        | 5.46              | 0.18           | [0.01, 0.35]          | 5.19          | 5.17              | -0.02          | [-0.15, 0.20]          | 0.23      |
| Trolling               | 5.24                        | 5.48              | 0.24           | [0.05, 0.42]          | 5.26          | 5.24              | -0.02          | [-0.13, 0.17]          | 0.31      |

# Appendix 3: All 18 fake news items participants viewed pre-post by badge.

Dr. Toe Lee afratTamica | Professor of He HED AND I Official Twitter account of Warren Büffet aWarrenBüffet | Billionsire, The 8th season of As a Scientist, I do not #GameOfThrones will be believe there is enough Investment advice: buy only postponed due to a salary evidence to suggest humans stocks that make you happy. dispute. can influence the climate. **Emotional Content** Daily Web Mews Animals News Parents Neskly NEWS ALERT: Baby formula Brutally beaten senior HEARTBREAKING story: baby desperately begs for help at hospital. is given TERRIBLE medical care only caused HORRIFIC outbreak of elephant gets horribly hurt new, terrifying disease after falling off a ledge, among helpless infants. mother elephant cries for after ENORMOUS wait. HOURS! Parents despair. **Conspiracy** Theories Daily Web News eDailywebrews | Daily updates on politics and more Row News at 1 plauleurAtt / Live action news on Live News Now eLiveNewsNow | News and opinion BREAKING: Insurance Scientists discovered The Bitcoin exchange rate solution to greenhouse effect years ago but aren't companies are using your is being manipulated by a phone to track your fast small group of rich allowed to publish it. report claims. food consumption. bankers, #InvestigateNow **Discrediting Opponents** ENG News and opinion. Susan P slusser 1 Retwort down't imply Che International Post Online einternationalgestonline ( Online Report reveals scientists' The mainstream media has Medical students only underlying agendas and receive a total of 5 hours of tutoring in nutrition. Don't trust doctors' been caught in so many lies that it can't be trusted as preferences, casting doubt on their findings. a reliable news source. #QuestionScience #FakeNews dietary advice. Polarisation Next Global The Daily Chronicle J Rapid Updates The myth of "equal IQ" New study shows right-wing Worldwide rise of left-wing between left-wing and people lie much more than extremist groups damaging world economy: UN report. right-wing people exposed. left-wing people. #TruthMatters Trolling Johnson Crude 011 Quad Media Joe Stephensohn Another shark loan for Hey eLeoDiCaprio, it's A sandwich maker at Subway developing countries snowing and freezing in New just took a bite out of my @WorldBank? York. Could use some of sub! Are you starving your #WorldOfExtortion that global warming you're employees @Subway? #HumanBanking always going on about!

#### Chapter 3

Appendix 4: All 18 chat screenshots (fake, real news items) participants viewed pre-post by badge.



Appendix 5: Appendix 1: Average reliability (pre-post) judgments for each fake news badge by experimental condition.

|              | Control          |                   |                   |                  |                   |                   |                        |      |
|--------------|------------------|-------------------|-------------------|------------------|-------------------|-------------------|------------------------|------|
|              | M <sub>pre</sub> | M <sub>post</sub> | M <sub>diff</sub> | M <sub>pre</sub> | M <sub>post</sub> | M <sub>diff</sub> | 95% CI <sub>diff</sub> | d    |
| Fake expert  | 2.6              | 1.8               | -0.8              | 2.7              | 2.5               | -0.2              | [0.4;0.6]              | 0.6  |
| Emotion      | 2.8              | 2.1               | -0.7              | 2.9              | 2.7               | -0.2              | [0.33;0.57]            | 0.5  |
| Polarisation | 2.6              | 1.9               | -0.7              | 2.7              | 2.4               | -0.3              | [0.2;0.46]             | 0.39 |
| Escalation   | 2.5              | 2.0               | -0.5              | 2.7              | 2.5               | -0.2              | [0.25;0.48]            | 0.43 |
| Real         | 2.3              | 2.0               | -0.3              | 2.4              | 2.3               | -0.1              | [0.1;0.3]              | 0.31 |

#### **Experimental condition**

## Appendix 6:Linear regression for demographics.

|  |          |      | 95% Confidence<br>Interval |       |        |           |  |
|--|----------|------|----------------------------|-------|--------|-----------|--|
| Predictor  | Estimate | SE   | Lower                      | Upper | -<br>t | р         |  |
| Intercept <sup>a</sup>   | -0.35    | 0.11 | -0.58                      | -0.13 | -3.14  | 0.00<br>2 |  |
| Gender:  |          |      |                            |       |        |           |  |
| Male – Female  | 0.08     | 0.04 | -0.002                     | 0.17  | 1.89   | 0.05<br>8 |  |
| Other, please specify: – Female  | -0.24    | 0.63 | -1.49                      | 0.9   | -0.38  | 0.69      |  |
| Age:   |          |      |                            |       |        |           |  |
| 25-34 - 18-24  | 0.017    | 0.08 | -0.149                     | 0.18  | 0.2    | 0.83      |  |
| 35-44 - 18-24  | -0.044   | 0.08 | -0.21                      | 0.12  | -0.52  | 0.60      |  |
| 45-54 - 18-24  | -0.007   | 0.09 | -0.18                      | 0.17  | -0.08  | 0.93      |  |
| 55 or older – 18-24  | -0.06    | 0.09 | -0.24                      | 0.11  | -0.67  | 0.50      |  |
| Education:   |          |      |                            |       |        |           |  |
| Degree or Graduate education (eg<br>BSc, BA) – A-level education   | 0.021    | 0.06 | -0.09                      | 0.14  | 0.34   | 0.73      |  |
| GCSE-level education – A-level education   | -0.007   | 0.07 | -0.156                     | 0.14  | -0.10  | 0.91      |  |
| Post-graduate education (eg PhD, MSc, MA) – A-level education  | -0.003   | 0.07 | -0.14                      | 0.14  | -0.04  | 0.96      |  |
| Undergraduate education (eg<br>University examinations but not<br>completed degree) – A-level<br>education | 0.014    | 0.07 | -0.13                      | 0.16  | 0.18   | 0.85      |  |
| Political Spectrum:  |          |      |                            |       |        |           |  |
| 2-1  | -0.10    | 0.09 | -0.28                      | 0.07  | -1.13  | 0.25      |  |
| 3 – 1  | -0.09    | 0.09 | -0.28                      | 0.08  | -1.07  | 0.28      |  |
| 4 – 1  | 0.05     | 0.08 | -0.11                      | 0.23  | 0.62   | 0.52      |  |
| 5 – 1  | -0.17    | 0.09 | -0.36                      | 0.02  | -1.75  | 0.08      |  |

| 6 – 1 | -0.04 | 0.12 | -0.28 | 0.19  | -0.34 | 0.73 |
|-------|-------|------|-------|-------|-------|------|
| 7 - 1 | -0.39 | 0.19 | -0.77 | -0.01 | -2.06 | 0.04 |

<sup>a</sup> Represents reference level

#### Chapter 4

All the necessary information needed to replicate the findings and methods, including our datasets, Qualtrics surveys, the full list of items (social media posts), preregistrations, supplementary tables, figures and analyses, and the analysis and visualisation scripts can be found on the Open Science Framework (OSF) page: <u>https://osf.io/mbqwj/</u>.

#### Chapter 5

#### Appendix 7: Attack message on fictitious chemical presented at T1.

Most people are under the impression that chemicals are vigorously tested before they are allowed on the market but this is simply not true. In fact, Tryptostine, the main ingredient used in most antibacterial products was suspiciously never vetted by the U.S. Food and Drug Administration (FDA) at all. Though the government has tried to keep this under the radar, a recently leaked report shows that it has taken the FDA almost 40 years to review it. Tryptostine is one of many drugs that governments approve without conducting barely any prior testing on its effectiveness and safety. Though its removal would be costly, many anecdotal stories have been posted online to expose the horrifying consequences of this chemical. However, these efforts keep being shut down and removed by social media platforms. So the government is not only complacently tolerating the risks of tryptostine, it actively shuts down any efforts to inform the public about it!

Disturbingly, tryptostine can be found in hand sanitisers, soaps and body washes, toothpastes, some cosmetics, in consumer products like clothes, kitchen utensils, and even in children's toys! Shockingly, companies don't even have to disclose the use of tryptostine in their products. Although many countries of the EU have immediately banned tryptostine after testing it, other countries remain suspiciously complacent and quiet on this issue! Recently, Avon and Johnson & Johnson have followed lead and announced to cut this dangerous chemical from its products but it still remains in many products we use every single day. A series of eye-opening investigations showed how the chemical causes malformations in mice and rats, a petition created by a group of concerned citizens, nurses, and practitioners went viral on Facebook. The organisers stress the "horrifying and concerning findings" shown in this report and call for immediate action. One of the studies in this report found traces of Tryptostine in human breast milk, urine, and blood plasma and experts stress that "direct exposure to Tryptostine can cause serious irritations and even antibiotic resistance". Professor Walsh,

from the Institute of Chemical Safety, states that "more evidence is needed to understand whether Tryptostine is carcinogenic, mutagenic or to what extent it could disturb the development of an embryo or fetus. To do so, the government must start taking this issue serious enough to fund research on such a controversial chemical".

#### Appendix 8: Inoculation treatment used at T1.

Most of us are confident in our ability to detect and resist misinformation and we wouldn't think about sharing it. Whether you are aware of it or not, there is a high likelihood that in the past you have fallen for and perhaps even shared misinformation with others. It is very likely that in this study you will come across misinformation that is so persuasive, that it could make you second-guess your existing beliefs towards the topic. Research shows that we overestimate our ability to spot and resist falsehoods. To stop harmful misinformation from going viral in the first place, you should equip yourself with the necessary skills to resist misinformation in the future. Below, you can learn about some of the most commonly used strategies used by media-manipulators and how to respond when you encounter them.

One way to make falsehoods go viral is using emotional language. Emotional content is not necessarily "fake" or "real" but rather deliberately plays into people's basic emotions such as fear, anger, and empathy (e.g. "Shocking results on the side-effects of cotton ear buds in this recently leaked report!"). Because the use of emotionally charged words provokes outrage, people tend to engage, react, and share manipulative content before critically assessing its accuracy. Thus, rather than reacting immediately, experts advise to approach inflammatory content with caution and a critical eye. Secondly, a lie can appear more reliable when a source backs it up. Which why the use of fake experts can be an easy and manipulative way to make content seem more credible. Even if the source doesn't exist or is being misquoted, it is easy to get distracted by fancy degrees and technical terminologies when viewing shocking information (e.g., 'Dr Isley from the University of Camford reports that "chemical used in toothpaste has adverse consequences on children's development"). Lastly, in times of uncertainty, such as during this pandemic, it's tempting to look for hidden motives or causes behind what is going on. Conspiracy theories often exploit such tendencies by picking a topic, a target to blame, and connecting dots to "prove" that the occurring events are orchestrated (e.g. "Doctors and nurses agree on dangers of 5G radiation but scientists refuse to publish their findings! What is the government covering up?"). While it is health to be sceptical, many things cannot simply be reduced to a single cause. Research suggests that being aware of these strategies can help you spot and resist manipulation so stay vigilant!

#### Appendix 9: Control message on fictitious chemical used at T1.

Chemicals are **essential building blocks** for everything in the world. All living matter, including people, animals and plants, consists of chemicals. All food is made up of chemical substances. Chemicals in food are largely harmless and often desirable – for example, nutrients such

as *carbohydrates*, *protein*, fat and fibre are composed of chemical compounds. Many of these occur naturally and contribute both to a rounded diet and to our eating experience. To give an example, Tryptostine is a compound that is naturally found in the earth's crust in a crystalline state. It can be obtained from mining and purifying quart.

It is present in a variety of plant foods, including vegetables and cereal grains (e.g. beets, sprouts, and oats). It is also found in animals and the human body as it is a component of human ligaments, cartilage and, musculature. Additionally, given its ability to block moisture absorption and prevent ingredients from clumping together, Tryptostine is used in food products to help retain their texture. It's most often found in granular or powder products, because as the European Food Safety Authority (EFSA) describes it, "it increases speed of dispersion, keeping the food particles separated and permitting the water to wet them individually instead of forming lumps.". This is why it is approved to be used as an additive to thinks like baking ingredients, protein powders, and dried spices. Tryptostine has a variety of uses in industries ranging from food and cosmetics to construction and electronics (e.g. cans, impermeable films, paints, and soil conditioners).

Given that chemicals can have a variety of toxicological properties, the safe use of chemicals such as triptrostine are closely monitored and informed by consistent scientific research. The evidence suggests that about 40 milligrams daily of tripostine from your diet may be linked with stronger bones. The U.S. Food and Drug Administration (FDA) has stated that tripostine added to food cannot exceed 2 percent of the food's total weight. The Expert Group on Vitamins and Minerals comply with the highest safety standards and identified a safe upper level for daily consumption of Tryptostine at 12 milligrams per kg body weight/day (for a 60 kg adult). The average individual has an average daily intake of 100 milligrams in total.

#### Appendix 10: Visualisation of the 'Degrees of Resistance' Index.



#### Appendix 11: Intercoder reliability analyses for content measures.

| Measures                   | Cohen's Kappa |
|----------------------------|---------------|
| Spreading inoculation (T1) | 0.71          |
| Spreading resistance (T1)  | 0.69          |

### Chapter 8: Appendices

| Spreading misinformation (T1) | 0.79 |
|-------------------------------|------|
| Resistance Index (T1)         | 0.69 |
| Spreading inoculation (T2)    | 0.81 |
| Spreading resistance (T2)     | 0.92 |
| Spreading misinformation (T2) | 0.88 |
| Resistance Index (T2)         | 0.72 |
|                               |      |

Appendix 12: PIT messages (T2) as the vicarious inoculation treatment.



Do not fall for misinformation about FDA non testing drugs! Don't share fake news, don't fall for it!

6:33 PM · Jan 24, 2022



•••

...

...

•••

...

People who are against Tryptosine use inflammatory and emotionally charged language to try and get you on their side.

6:40 PM · Jun 18, 2021



Fake news uses fake authority, emotions and conspiracies, so if you read 1 article about a chemical that uses all that, maybe don't trust it

1:42 PM · Mar 24, 2022



Some people have been spreading information and misinformation about a chemical that does not exist. Did you fall for it?

7:45 PM · Feb 28, 2022



misinformation uses emotional words, the opinion of fake experts and conspiracy theories. what they say about tryptostine is false

11:45 AM · Feb 27, 2022



Someone made up a fake chemical and wrote whole articles about it! Fake news are getting dank

5:13 PM · Feb 24, 2022



Be careful what you read about inflammatory articles about Tryptosine. The lack of scientific studies make this a conspiracy theorist game.

9:52 AM - Feb 13, 2022

Appendix 13: PIT messages (T2) that scored lowest on the Resistance Index and were used instead of a traditional attack message.



7:14 PM · Mar 16, 2022

Appendix 14: Fictitious, neutral, and unrelated treatment for the control condition.



Hands down, the best political message to all Russians so far comes from Arnold Schwarzenegger!  $\P \subseteq$ .

7:13 PM · Mar 29, 2022



Just watched this Ted talk that says the best predictor of success isn't IQ or even good looks. It's grit. Highly recommend watching it

4:10 PM · Jan 19, 2022



Can't believe Deptford has been selected as the best up and coming area in London. Can we stop with all this gentrification PLEASE????

11:02 AM · Feb 15, 2022



Over in the US - a shopping mall has installed its first 'RoboBurger' outlet. I can't believe fast food is now being served via vending machine...

6:47 PM - Jan 29, 2022



...

...

....

...

...

everyone's talking about will smith slapping chris rock at the oscars but i can't believe no one's talking about the fact that jason derulo FELL DOWN THE STAIRS-

12:22 PM · Feb 23, 2022



South Africa's unemployment rate hits new record high in Q4 2021

12:22 PM · Feb 23, 2022



•••

...

In China's Wall Street, bankers and traders sleep in offices to beat Shanghai COVID lockdown

3:34 PM · Feb 27, 2022