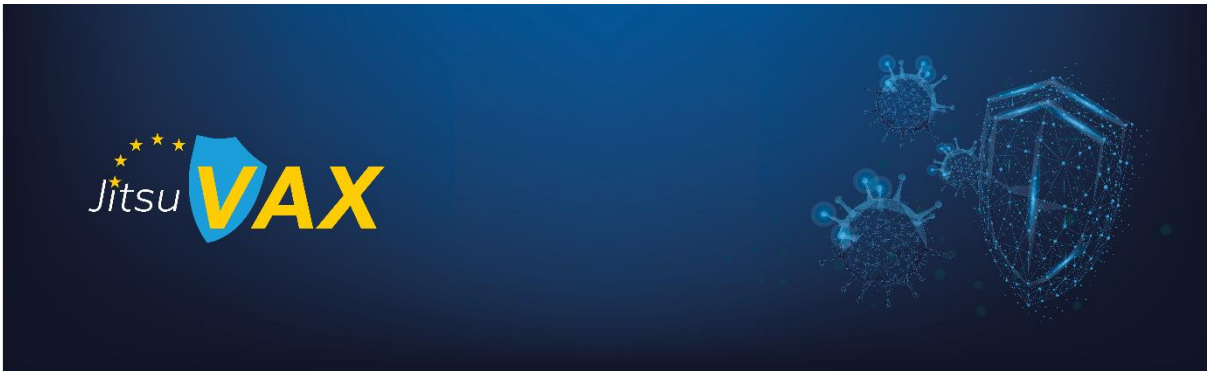

JITSUVAX:
Jiu-Jitsu with Misinformation in the Age of Covid

Postinoculation talk on social media

September 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964728 (JITSUVAX)

Co-funded by the Horizon 2020 programme of the European Union

JITSUVAX Deliverable 3.3
Postinoculation talk on social media

Project title:	JITSUVAX: Jiu-Jitsu with Misinformation in the Age of Covid
Grant agreement:	964728
Duration:	April 2021-March 2025
Website:	https://sks.to/jitsuvax
Coordinator:	Stephan Lewandowsky
Deliverable number:	3.3
Deliverable Title:	Post-inoculation talk on social media
Dissemination level:	Public
Version:	1
Authors:	Jon Roozenbeek, Sander van der Linden
Reviewed by:	Stephan Lewandowsky
Contacts:	sv395@cam.ac.uk, jitsuvax@bristol.ac.uk

Consortium: **University of Bristol**, Beacon House Queens Road, Bristol, BS8 1QU, UK
Universität Erfurt, Nordhauser Strasse 63, Erfurt 99089, Germany
The Chancellor Masters and Scholars of the University of Cambridge, Trinity Lane, The Old Schools, Cambridge, CB2 1TN, UK
Turun yliopisto, Yliopistonmaki, Turku 20014, Finland
Observatoire Regional de la Sante,27 Boulevard Jean Moulin, Marseille 13005, France
Universidade de Coimbra, Paço das Escolas, Coimbra 3001 451, Portugal.

Contents

Summary

Psychological "inoculation" is a largely preemptive approach to building resilience to misinformation. In the same way that a vaccination stimulates the body into generating antibodies by imitating an infection, which can then fight the real disease when an actual infection occurs, psychological inoculation stimulates the generation of counter-arguments that prevent subsequent misinformation from sticking (McGuire, 1964). As van der Linden et al. (2017) put it: "By preemptively warning people against misleading tactics and by exposing people to a weakened version of the misinformation, cognitive resistance can be conferred against a range of falsehoods in diverse domains" (p. 1141).

Over the last 50 years, inoculation theory has demonstrated its effectiveness as a strategy to confer psychological resistance against unwanted persuasion (Banas & Rains, 2010). Yet only recently, research has explored the possibility of a "broad-spectrum vaccine" against misinformation, i.e. rather than focusing on inoculating against a specific falsehood, the aim is to inoculate people against the larger techniques of disinformation (Basol et al., 2020; Cook et al., 2017; Roozenbeek & van der Linden, 2019). Specifically, interventions have focused on pre-emptively exposing participants to - and subsequently refuting - more generic techniques of misinformation, using an interactive online game known as *Bad News* (Roozenbeek & van der Linden, 2019).

In *Bad New*s, players start out as an anonymous "netizen" and they eventually rise to manage their own fake news empire. The game simulates a social media feed and exposes the player to weakened doses of six common misinformation techniques (e.g., conspiracy theories, fearmongering, fake experts, polarization) in an attempt to cultivate cognitive antibodies. Yet, although these games have been played by millions of people around the world and shown to help people recognize misinformation (Basol et al., 2020; Roozenbeek & van der Linden, 2019; Maertens et al., 2021), little remains known about the effects of "postinoculation talk"; that is, what people spontaneously talk about following the inoculation (such as in public discussion forums on social media). The notion of postinoculation talk is crucial but it allows for the possibility of vicariously passing on the inoculation within a social network, which could lead to herd immunity (Pilditch et al., 2022). This study offers the first exploration of postinoculation talk "in the wild".

A crucial precondition for an analysis of postinoculation talk is the existence of a large body of text generated by social interaction after an inoculation intervention has occurred. Clearly, unless a large group of people is exposed to an inoculation and then decides to hold a conversation about it, no analysis can take place. This creates unique constraints for the present research because the availability of text to analyze is not under the experimenters' control.

We were able to successfully adapt to those constraints, but it necessitated a deviation from our original plans. Specifically, although we developed two inoculation games (*BadVaxx* and *VaxBN*) as part of the JITSUVAX project (WP2.2), and although the development of games was completed nearly on time, a delay in their public roll-out prevented us from collecting inoculation talk for the new games. We therefore decided to analyse a large body of relevant text that was gathered from a previous game, known as *Bad News*.

The *Bad News* game is a direct predecessor to the *Bad Vaxx* and *VaxBN* games, and works on a nearly identical premise and game structure. The game covers vaccine misinformation in one of its 6 levels, alongside a range of other topics such as climate change and COVID-19. It was also featured on a large reddit thread which gained 67,000 upvotes in 2019. Therefore, the decision was made to use the data from that existing thread for the postinoculation talk analysis. The results reported here were thus obtained in the planned manner and in the planned theoretical and conceptual context – the only unplanned difference is that the topic domain is not specific to vaccinations.

Using topic modelling and sentiment analysis dictionaries, we analyze comments (*N* = 36,127) by Reddit users who posted on a thread about the *Bad News* game, investigating 1) how social media users reacted to learning about inoculation interventions, and 2) to what extent users engaged in issue-relevant postinoculation talk We find indications that Reddit users who commented on the *Bad News* thread subsequently engaged more in independent postinoculation talk and counterarguing, including on the topic of vaccination, compared to a control group. We also find that social media environments can facilitate what we call "*meta-inoculation talk*": discussions about the tenets and societal relevance of inoculation theory itself.

Scope and purpose of this document

This document reports on a study conducted on Reddit as part of WP2.2, the purpose of which was to investigate the extent to which playing an inoculation game about misinformation subsequently leads to so-called "postinoculation talk", i.e., people engaging in discussions and counterarguing against the misinformation against which they have been inoculated. This document lays out the background, methodology, results, and other findings of this study.

Project overview

Vaccine hesitancy—the delay or refusal of vaccination without medical indication—has been cited as a serious threat to global health by the World Health Organization (WHO), attributing it to misinformation on the internet. The WHO has also identified Health Care Professionals (HCPs) as the most trusted influencers of vaccination decisions.

JITSUVAX will leverage those insights to turn toxic misinformation into a potential asset based on two premises:

1. The best way to acquire knowledge and to combat misperceptions is by employing misinformation itself, either in weakened doses as a cognitive "vaccine", or through thorough analysis of misinformation during "refutational learning".
2. HCPs form the critical link between vaccination policies and vaccine uptake.

The principal objective of JITSUVAX is to leverage misinformation about vaccinations into an opportunity by training HCPs through inoculation and refutational learning, thereby neutralizing misinformation among HCPs and enabling them to communicate more effectively with patients. We will disseminate and leverage our new knowledge for global impact through the team's contacts and previous collaborations with WHO and UNICEF.

Background

Over the last 60 years, inoculation theory, often referred to as the "grandparent theory of resistance to attitude change" (Eagly and Chaiken, 1993, p. 561), has proven its effectiveness as a messaging strategy to build psychological resistance against unwanted persuasion (McGuire and Papageorgis, 1961b; Banas and Rains, 2010; Compton, 2013). Psychological inoculations, much like medical vaccinations, expose an individual to a weakened version of a particular pathogen (or malicious persuasion attempt), which triggers a protective response in the form of (mental) antibodies. These processes eventually lead to greater psychological resistance against subsequent persuasive attacks (Ivanov *et al.*, 2012; Compton, 2013).

Since its original formulation by McGuire in the 1960s (McGuire and Papageorgis, 1961b; Papageorgis and McGuire, 1961; McGuire, 1964), scholars have tested inoculation theory in a variety of issue domains, including health (Parker, Ivanov and Compton, 2012; Compton, Jackson and Dimmock, 2016; Ivanov, 2017) and politics (Compton and Ivanov, 2013). More recently, inoculation interventions have

also been applied to strengthen resistance against conspiracy theories about 9/11 and vaccinations (Banas and Miller, 2013; Jolley and Douglas, 2017). Crucially, inoculation theory has also been shown to be effective against online misinformation and "fake news", including important political issues such as climate change (van der Linden, Leiserowitz, *et al.*, 2017; Maertens, Anseel and van der Linden, 2020), online extremism (Saleh *et al.*, 2021) and COVID-19 misinformation (Basol *et al.*, 2021).

Within the domain of online misinformation, inoculation research has recently seen three significant advances: 1) a shift in focus from inoculating people against specific misinformation to interventions that inoculate individuals against manipulation *techniques* (e.g., Cook et al., 2017; van der Linden & Roozenbeek, 2020); 2) a move away from passive (e.g., through reading) towards *active* inoculation, in which individuals invest a significant amount of cognitive effort in the inoculation process (McGuire and Papageorgis, 1961a), for example, by playing an interactive perspective-taking game where they actively generate their own content (Roozenbeek and van der Linden, 2018) and 3) a shift from viewing inoculation as a mostly *intrapersonal* process to also include interpersonal communication such as postinoculation talk (Compton and Pfau, 2009; Dillingham and Ivanov, 2016; Rains, 2018).

These advances have led to new theoretical developments such as distinguishing between therapeutic and purely prophylactic inoculations (Compton, 2019) and have been beneficial in improving the scalability of inoculation interventions by broadening the scope of the cognitive "vaccine" (Roozenbeek and van der Linden, 2019). Less is known, however, about the conditions under which this occurs; studying the diffusion process of messages in a social media context is one of the ways in which this research advances not only the study of postinoculation talk, but also interpersonal communication on the internet more generally.

Active inoculation through gamification

As part of JITSUVAX (WP2.2), we developed two gamified inoculation interventions that build psychological resistance against vaccine misinformation: *VaxBN* and *Bad Vaxx*. The target audiences for these games were health care providers (HCPs) and the public at large, respectively. In the original proposal, one game was planned. However, as the work developed it became apparent that a second game, specifically targetting HCP's, would be a useful addition to the JITSUVAX output. Therefore, we developed both Bad Vaxx for the general public and VaxBN for the HCP audience. As a result, testing and public launches of the games have been delayed. Their full development and testing will be described in Deliverable 2.2, due in month 24.

For the reasons noted at the outset, we used an existing large thread on the subreddit r/science about the *Bad News* game for the present analysis. *Bad News* is a gamified inoculation intervention highly similar in scope to *Bad Vaxx* and *VaxBN* and developed by the same research team (UCAM). This thread gained a significant amount of upvotes (around 67,000) and comments (around 800). This thread therefore provided an excellent opportunity to explore postinoculation talk "in the wild", in line with the requirements of WP3.2.

In *Bad News*, players take the perspective a fake news creator (see Figure 1 for a screenshot).[1] The game has been played by over a million people and has been translated into more than 20 languages (Roozenbeek, van der Linden and Nygren, 2020).

[1] The game is free and can be played in any browser at www.getbadnews.com.

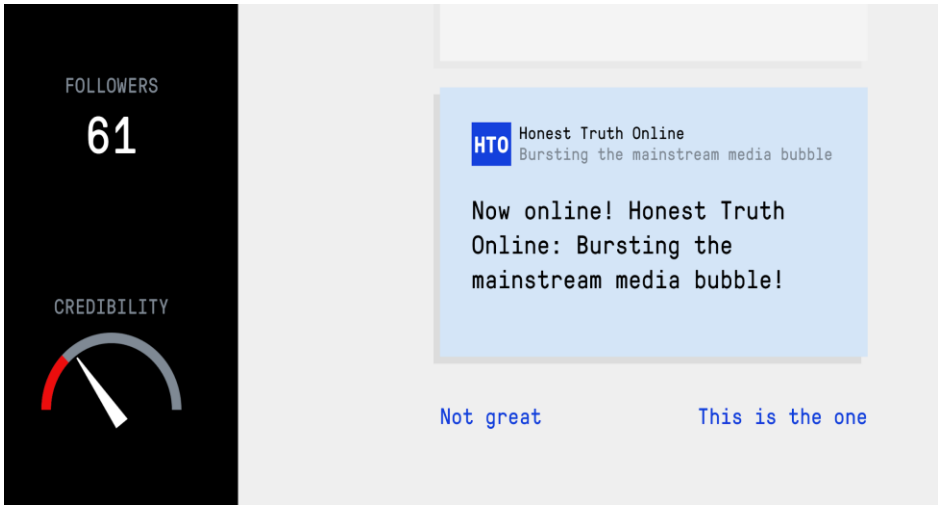

Figure 1. A screenshot of the *Bad News* game environment (www.getbadnews.com).

The game simulates a social media environment and is choice-based with a simple interface: players are tasked with gaining followers and building credibility for their fake news website. If their credibility reaches 0, they lose. Over the course of the game, players earn six badges, one per common online misinformation technique: impersonating news producers and fake accounts online (Reznik, 2013; Goga, Venkatadri and Gummadi, 2015; BBC News, 2018), using emotionally charged language (Brady *et al.*, 2017; Crockett, 2017; Berriche and Altay, 2020), polarizing audiences by exploiting wedge issues (Rojecki and Meraz, 2016; Iyengar and Massey, 2018), spreading conspiracy theories (Lewandowsky, Oberauer and Gignac, 2013; van der Linden, 2013), discrediting opponents, for example through ad-hominem attacks (Walton, 1998; Lischka, 2017), and trolling people online to evoke an excessive response (Griffiths, 2014) (see Roozenbeek & van der Linden, 2019 and van der Linden & Roozenbeek, 2020 for a detailed overview of these techniques). Throughout the game, players gradually grow from being an anonymous social media user to a successful misinformation tycoon. Various topics are covered, but vaccine misinformation is specifically addressed in the *Conspiracy* scenario.

During the approximately 15 minutes of gameplay, players are forewarned and exposed to a weakened dose of misinformation techniques through a combination of perspective-taking and active experiential learning. Here, the "bad guy" perspective serves as the "motivational threat" component (compelling people to defend their attitudes against attacks) of the inoculation treatment (Compton and Pfau, 2005; Compton, 2013; Richards and Banas, 2018). Subsequent research has shown that the *Bad News* intervention is effective at improving people's ability to spot misinformation (Roozenbeek and van der Linden, 2019), increases people's confidence in their ability to discern misleading information (Basol, Roozenbeek and van der Linden, 2020), shows similar results in five different cultural contexts and language versions of the game (Roozenbeek, van der Linden and Nygren, 2020), and can confer detectable inoculation effects for at least 13 weeks after initial gameplay if players are periodically given brief reminders or "booster shots" (Maertens *et al.*, 2021). We refer to Roozenbeek, Maertens, McClanahan, & van der Linden (2021) for a detailed analysis of the game, its items, and test performance.

Post-inoculation talk online

> *Inoculation research to date has stressed what goes on in people's minds during the process of resistance. What inoculation research needs to do next is to learn what goes on in people's discussions and dialogues with others following the administration of inoculation treatments* (Compton and Pfau, 2009, p. 21).

The present study advances the postinoculation talk and communications literature by answering this call: documenting online discussions following an active inoculation treatment. For most of its history,

inoculation theory has focused on intrapersonal processes such as threat and so-called "subvocal" counterarguing, whereas it remains relatively less clear if and what people talk about following a successful inoculation (Ivanov *et al.*, 2015). This is important as postinoculation talk has been referred to as the metaphorical "syringe" injecting the immunization from one individual to the next through a process of social diffusion (Compton and Pfau, 2009). Several open questions about the effectiveness of active inoculation interventions remain, particularly with regard to what happens when people spread the cognitive "vaccine" (e.g., resistance acquired while playing the *Bad News* game) through counterarguing, in what has become known as "postinoculation talk" (Compton and Pfau, 2009; Ivanov *et al.*, 2012; Dillingham and Ivanov, 2016).

Building off of the idea that inoculation messages can both unsettle one's confidence and at the same time motivate advocacy (Compton and Pfau, 2009)—and that these two responses often lead to talk—scholars have explored inoculation theory's effect on continuing talk following an inoculation treatment. Previous research has shown that inoculation treatments increase inoculated individuals' talk about the target issue and that this talk appears to bolster resistance to future persuasion attacks (Ivanov *et al.*, 2012, 2015), as well as issue-relevant advocacy attempts, i.e., counterarguing (Ivanov *et al.*, 2015). Furthermore, postinoculation talk has been shown to strengthen belief certainty among inoculated individuals (Dillingham and Ivanov, 2016) and shore up behavioral intentions that are consistent with the inoculated position (Ivanov *et al.*, 2017). However, the available body of research on postinoculation talk remains relatively small, particularly with regards to active inoculation interventions. In particular, little evidence is available about how people interact with inoculation interventions "in the wild", and, crucially, to what extent inoculated individuals engage in postinoculation talk with others. This is important because as Compton and Pfau (2009) noted, outside of controlled laboratory settings, people may exhibit discussion that could both facilitate and hinder the inoculation process.

Most recently, Rains (2018) suggested looking at the group-level implications of inoculation using big data research methods such as web-scraping, social network analysis, and topic modelling. Addressing these questions is crucial for developing further insight into the boundary conditions of inoculation theory, such as the potential for achieving "herd immunity" against manipulation attempts via postinoculation talk in social networks (Compton and Pfau, 2009; van der Linden, Maibach, *et al.*, 2017). In particular, important open questions remain about the extent to which hearing about the inoculation treatment from others (online)—but not actually going through the inoculation treatment itself—induces postinoculation talk and counterarguing.

The present study

This study is the first to explore the above questions by examining postinoculation talk about a gamified inoculation intervention on the social media platform Reddit.[2] Reddit is an open forum site, with a large number of subforums or "subreddits" dedicated to a variety of topics ranging from video games to politics and memes, on which users can post in as many different subreddits as they want (Roozenbeek and Salvador Palau, 2017). Reddit has over one million online forums and about 330 million active monthly users (Klein, Clutton and Dunn, 2019).

In June of 2019, a submission on the r/science subreddit about the *Bad News* game and a research article that was published about its effectiveness (Roozenbeek and van der Linden, 2019) went viral,

[2] All analyses and web-scraping scripts, as well as the complete dataset, are available on the OSF: https://osf.io/4q6sh/?view_only=96cdf1571e6844d9a6a56ca8c95a502a (peer review link).

gaining over 67,000 upvotes and 1,900 unique comments in a single day (Reddit, 2019a).[3] This thread provided sufficient data to analyze the proliferation of postinoculation talk "in the field", by analyzing the language use of Reddit users who posted on this thread. Specifically, we investigate the following research questions:

1) How did Reddit users react to the viral *Bad News* inoculation intervention and learning about inoculation theory more generally?

2) To what extent did Reddit users who posted on the *Bad News* Reddit thread subsequently engage in issue-specific postinoculation talk on Reddit?

With respect to the second question, we conceptualize engaging in active discussion about inoculation theory within a particular context, such as occurred in the *Bad News* thread, as a form of active inoculation that may induce postinoculation talk: participants in the threads did not merely hear about *Bad News,* or comment on the game or inoculation theory, but also invested significant cognitive effort into discussing misinformation techniques and inoculation theory itself.

Rather than a randomized experimental design, we use big data methods to analyze postinoculation talk on social media (Rains, 2018; Fong *et al.*, 2021). Specifically, we follow the standard methodology for research conducted with Reddit data, using a case control study design with two treatment groups and a control group (Klein, Clutton and Dunn, 2019). The treatment groups consist of posts by Reddit users who posted in the *Bad News* threads. Because merely commenting on the *Bad News* Reddit thread is not the same as playing through the entire game (as only the latter constitutes going through the inoculation treatment proper), we distinguish between users who we can be reasonably sure played through the entire *Bad News* game (e.g., by posting their final score in the *Bad News* thread; the "inoculation" group) and users who posted in the thread but of whom we cannot know if they played *Bad News* all the way through (the "weak interaction" group).

The control group consists of posts by Reddit users who posted on another thread in the r/science subreddit about psychological research around the same time as the treatment thread, with similar popularity in terms of upvotes and comments (and hence can be said to share similar interests with users who posted in the *Bad News* thread), but who did not post in the *Bad News* thread itself. Concretely, the control group contains posts by Reddit users who commented on a r/science thread posted on June 14, 2019, which covered the publication of an article about the psychological benefits of spending time in nature (Reddit, 2019b; White *et al.*, 2019). Including this control allows us to address whether postinoculation talk about the topics relevant to the inoculation intervention discussed in this study is induced after posting on *any* Reddit thread about psychological research (as opposed to the *Bad News* thread specifically). With the above in mind, we arrive at the following hypothesis:

H1: Reddit users who posted on the *Bad News* thread engage significantly more in issue-relevant postinoculation talk on Reddit (i.e., about misinformation) compared to a control group.

Methodology

Sample and Procedure

Using PRAW, a Reddit API wrapper for Python, we first scraped all comments on the r/science Reddit thread about the *Bad News* game, along with posters' usernames. As one of the authors of the paper that was the subject of the *Bad News* thread (Roozenbeek and van der Linden, 2019) was asked to

[3] This subreddit (www.reddit.com/r/science) is a 'place to share and discuss new scientific research', according to the page description.

participate in an ad hoc AMA (ask-me-anything) in the thread, their comments were removed, as well as one post in the thread which only summarized the paper's abstract.

Next, we used the Python-based Requests library to obtain all Reddit posts (across all of Reddit) by users who commented on the *Bad News* Reddit thread, posted during the week after the thread was posted at 1:27pm UTC on June 26, 2019. We also obtained each post's subreddit, its karma (the number of upvotes), and the time elapsed (in days, from 1 to 7, one for each day of the week) from the moment the relevant thread was posted. For analysis purposes (see the "Results" section), aside from the full ("combined") sample, we separate this sample into two subgroups: "inoculated" Reddit users (who we can be reasonably sure played through the entire *Bad News* game, for example if they posted their final score in the *Bad News* thread) and "weak interaction" users who posted in the *Bad News* thread but may not have played through the game .

For the control group, we followed the same procedure for users who commented on the control thread (obtaining all Reddit posts by these users during the week after June 26, 2019). We chose a time period of one week because recent research has shown that inoculation effects conferred by active inoculation treatments such as the *Bad News* game remain detectable for at least seven days after initial exposure (Maertens *et al.*, 2021). Table 1 shows the dataset details.[4]

Sample	***Time period***	***No. of unique users***	***No. of posts***
Bad News thread	26 June 2019	566	811
Inoculation group	26 June – 3 July 2019	35	1,181
Weak interaction group	26 June – 3 July 2019	531	20,045
Combined dataset	26 June – 3 July 2019	566	21,226
Control group	26 June – 3 July 2019	539	14,901
Total		**1,105**	**36,127**

Table 1. Dataset.

Method of analysis: topic modelling and Empath

To answer research question 1 and to inform the scope of research question 2, we make use of topic modelling using Latent Dirichlet Allocation or LDA (Blei, Ng and Jordan, 2003), in order to infer the dominant topics of discussion on the *Bad News* thread. Topic models are unsupervised machine learning classification algorithms, capable of grouping together semantically related words into topics in an otherwise unstructured corpus. LDA, specifically, assumes that each document is made up of a small number of topics, and that topics contain a small number of frequently used words. The topic probability distribution is assumed to have a sparse Dirichlet prior (also called a multivariate beta distribution). Topics are defined as clusters of words that occur frequently together. For example, a topic model might cluster the words "ball", "homerun", and "base" together, after which the topic "baseball" would be manually assigned to this topic. Topic modelling has seen a surge in popularity within the social sciences in recent years (Navarro-Colorado, 2018; Roozenbeek, 2020), and has previously been used to analyze language use on Reddit (Klein, Clutton and Polito, 2018; Klein, Clutton and Dunn, 2019). To build the model, we used the Gensim Python wrapper for the MALLET topic

[4] All four samples are similarly distributed in terms of the frequency of posts by each user in the dataset. Each sample has a positive skewness (>0.704), indicating that the majority of users in each dataset is responsible for a small number of posts. Because this is the case for every sample, we deem it possible to compare results between groups. See Table S6, Figure S2 and the "User post frequencies" tab in the dataset on the OSF.

modelling package (McCallum, 2002). For the construction, validation, and visualization of the topic model, we made use of Prabhakaran's workflow and the Matplotlib visualization library (2018, 2019), with minor adaptations. The Python scripts can be found on the OSF: https://osf.io/4q6sh/?view_only=96cdf1571e6844d9a6a56ca8c95a502a.

To analyze language use in Reddit posts (research question 2), we use Empath (Fast, Chen and Bernstein, 2016), an open source Python library capable of extracting linguistic characteristics from written text via dictionary categories, similar to and highly correlated with the commonly used LIWC or Linguistic Inquiry Word Count (Pennebaker *et al.*, 2015; Fast, Chen and Bernstein, 2016). Unlike LIWC (which is exclusively human-validated), Empath uses deep learning to capture specific words in a neural embedding, which learns associations between words and their context based on a test corpus of 1.8 billion words. Using this neural embedding, Empath's developers used similarity comparisons (most notably cosine similarity) between vectors (i.e., words) in a vector space (i.e., the "context") to map a total of 59,690 words onto 200 lexical categories, which were subsequently validated by humans. Using this approach, Empath is capable of analyzing written language using 200 pre-validated lexical categories, ranging from affective indicators such as *emotion*, *disgust*, and *disappointment* to more concrete categories such as *science*, *college*, and *clothing*.

In addition, Empath allows for the creation of custom categories based on one or more seed words (such as "misinformation" or "inoculation"). To do so, Empath queries a vector space model (VSM), trained by a neural network on a large corpus of text, in order to analyze the similarity between words across different dimensions of meaning and relevance (Fast, Chen and Bernstein, 2016). Importantly within the context of our study, Empath has the option to choose which corpus to use to generate the new lexical category from. One of these options is a corpus of Reddit posts, which we use here to generate custom Empath categories relevant to misinformation, inoculation theory, and postinoculation talk (such as counterarguing), as Empath does not have pre-defined categories that relate to these topics.

Units of text (in our case Reddit posts) are assigned a score for each category based on whether words in the text fall under that category. For example, if a Reddit post contains the word "newsworthy", and this word falls under the Empath category *journalism*, then this post is assigned a score for the *journalism* category. The category score for each post is based on weighted word frequencies across members of the category, normalized for the length of the post. In other words, the category score for a particular post is based on how many words from a particular category are used, divided by the total number of words in the post, so that the proportion of category-relevant words to total post length determines the category score (as opposed to the total number of category-relevant words). Empath's categorization scheme was partly trained on Reddit data, making it particularly useful for purposes of this study (Klein, Clutton and Dunn, 2019). We refer to the Methods Supplement for more details about Empath's category creation procedure.

Our next step is to determine which Empath categories to include in our analysis. Doing so requires a working definition of what constitutes "issue-specific postinoculation talk" in the context of pre-defined and custom-made Empath categories. Determining what lexical categories to include for the *Bad News* thread is not straightforward, because the *Bad News* game is about misinformation and news media in a general sense. To avoid including categories arbitrarily, we will therefore use the results from the topic model, described above, to inform which Empath categories are the most directly applicable to the topics of discussion in the *Bad News* thread (see the "Results" section).

Results

Discussions on the *Bad News* Reddit thread

In this section, we address research question 1, and discuss Reddit users' participation in the *Bad News* thread to investigate how users reacted to inoculation interventions and inoculation theory in an environment where talking about inoculation theory is explicitly facilitated. The results presented below will also inform what constitutes issue-specific postinoculation talk within the context of the *Bad News* game and its accompanying Reddit thread, and what Empath categories will be used to further analyze Reddit users' language use on the platform (research question 2).

First, to determine the main topics of discussion on the *Bad News* thread, we constructed a topic model with six topics over the 811 Reddit comments from the *Bad News* thread.[5] Figure 2 shows the results; each bar graph represents one of the six topics, with the words in each topic ranked by their in-topic weight. Dark colored bars represent a word's weight (or importance) within the topic whereas lighter colored bars represent their absolute frequency in the corpus.

Other comments by Reddit users who posted on the thread include *"Awesome scary game"*, and *"So in TL;DR, this whole website is basically [Sun Tzu's] Art of War but with media"*. The game also elicited threat (a key component of inoculation treatments; see Banas & Richards 2017) in other ways. For example, several (but not many) users expressed concern over being exposed to misinformation through the game: *"I find the implications of this game a bit scary. It immediately tries to goad its players into believing that anything claiming to be "against mainstream media" is automatically untrustworthy."*, and *"Couldn't this also be used to indoctrinate people?"*. Similar concerns have been raised about more "conventional" inoculation interventions, for example the possibility of inoculation against inoculation (Banas and Miller, 2013).

The remaining topics in Figure 2 are not about Reddit users' perceptions of the *Bad News* game per se. Topic 2 is about news media and journalism, and includes words relating more to the content of the game scenarios such as "source", "claim", and "bias". One discussion in the thread covered the veracity of several well-known conspiracy theories such as the 1964 Gulf of Tonkin incident and the sinking of the USS Maine in 1898. Topic 4 covers the methodological aspects of the study associated with the *Bad News* game, including inoculation theory (Roozenbeek and van der Linden, 2019) (with words such as "method" and "theory"; e.g., *"Inoculation theory? When regarding a task surely it's just called training?"*), but also misinformation specifically (the word "propaganda" appears in the topic, e.g., *"What do we do about the people who think all news is fake and propaganda? Would this [game] help them too?"*. Topics 3, 5, and 6 are somewhat similar in the sense that they all cover media literacy, education and critical thinking, with words like "teacher", "teach", "study", "learn', "skill", and "critical thinking" standing out. Topic 3 also covers the importance of science (e.g., "science", "study", and "knowledge", e.g., *"You would think education is the answer. But, there are other factors suggesting that education also needs critical thinking, fundamental science understanding, and sound logic to come to more accurate conclusions"*).

[5] The coherence score for this topic model is 0.43; see Kapadia (2019) and Prabhakaran (2018).

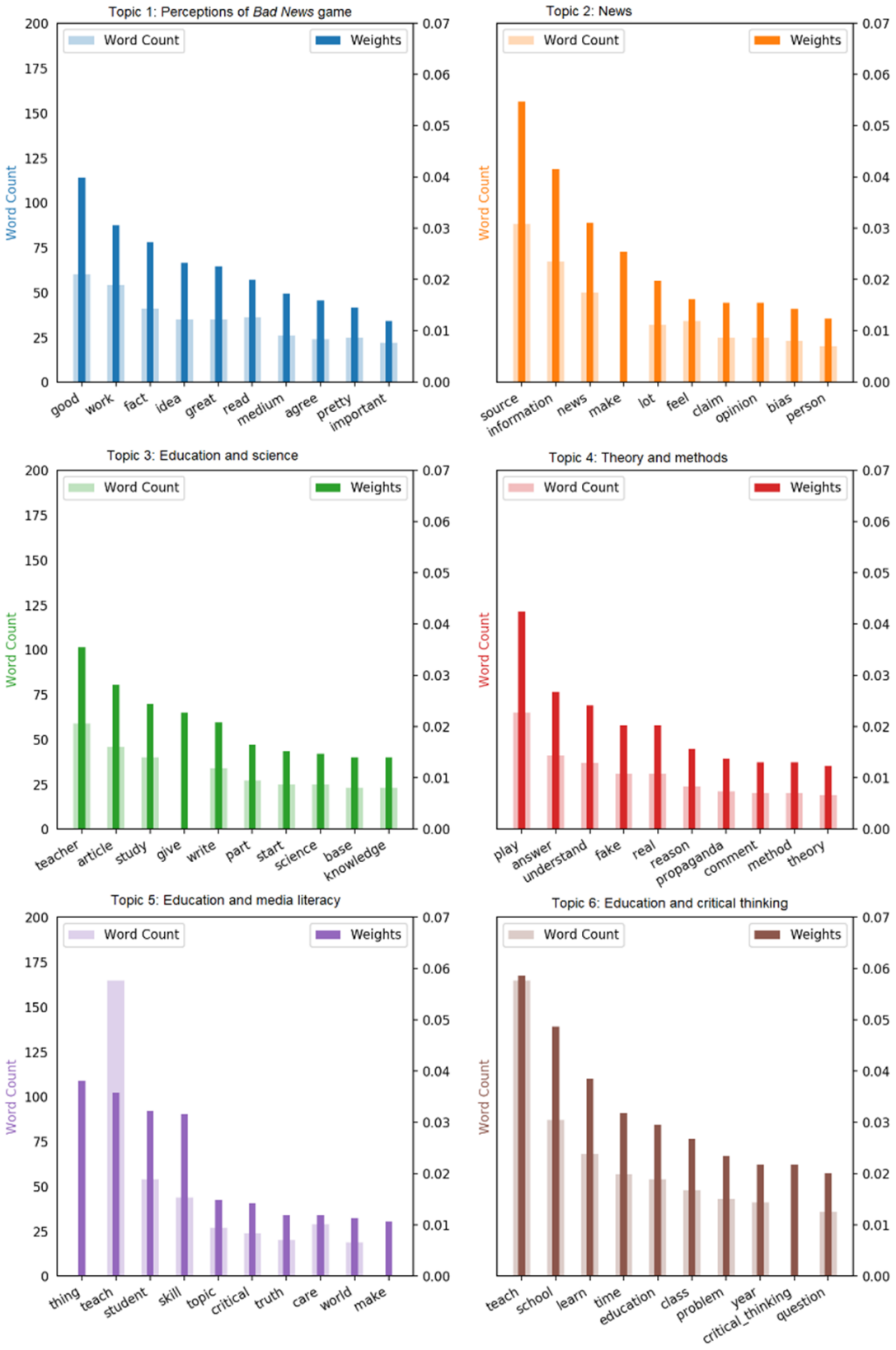

Figure 2. Visualization of 6-topic LDA model (MALLET) for the *Bad News* Reddit thread. Left *Y*-axis shows the word count. Right *Y*-axis shows the weight of each word in the topic.

To summarize, the topic model visualized in Figure 2 allows us to infer four dominant themes in the *Bad News* thread: science (including inoculation theory), (social) media, misinformation (including conspiracies), and education. These four themes inform what Empath categories are most directly relevant when it comes to issue-specific postinoculation talk, which is the subject of the next section. We therefore include the following pre-defined Empath categories: *science* (for the "science" theme), *journalism* and *social_media* (for the "(social) media" theme), and *school* (for the "education" theme). In addition, because Empath does not have pre-defined categories that relate specifically to misinformation, inoculation, and counterarguing, we include the following custom-made categories: *fake news*, *misinformation,* and *conspiracy,* all of which fall under the "misinformation" theme, as well as *counter_misinfo*, to cover the topic of countering misinformation (i.e., counterarguing, a core component of postinoculation talk). Finally, we also include two custom categories that relate to the "vaccination" metaphor that underlies inoculation theory: *vaccination* and *inoculation*.

Postinoculation talk on Reddit

This section explores issue-specific postinoculation talk (in the form of language use related to the Empath categories mentioned at the end of the previous section) by Reddit users during the week after posting on the *Bad News* thread. For purposes of clarity and brevity, we report the results for the combined *Bad News* sample; when separating this sample into the "inoculation" and "weak interaction" subgroups, the results are directionally similar, with minor differences between the "inoculation" and "weak interaction" groups; broadly speaking, the "inoculated" users appear to have engaged more in postinoculation talk than the "weak interaction" group, although differences are not significant due to the large difference between groups in the number of users and posts (see Table 1); see Tables S2 (Welch's ANOVA table) and S3 (Games-Howell post hoc tests).

To determine whether users who commented on the *Bad News* Reddit thread engage significantly more in issue-specific postinoculation talk than the control group, we conduct a series of Welch's *t*-tests (we do not conduct Fisher's independent samples *t*-tests because of unequal variances, see Table S6 and Figure S2)[6] on the raw number of posts, with condition (*Bad News* vs control) as the grouping variable and the normalized Empath scores for each category as the dependent variables. Table 2 and Figure 2 show the results. For the reader's convenience, we have also plotted the results in a single bar plot, for which we refer to Figure 3.

[6] We also conducted Mann-Whitney U-tests for each outcome variable; these broadly give similar results (albeit with different significance levels), although the Mann-Whitney U-tests are also significant for the *fake news, inoculation* and *science* categories. See Table S1.

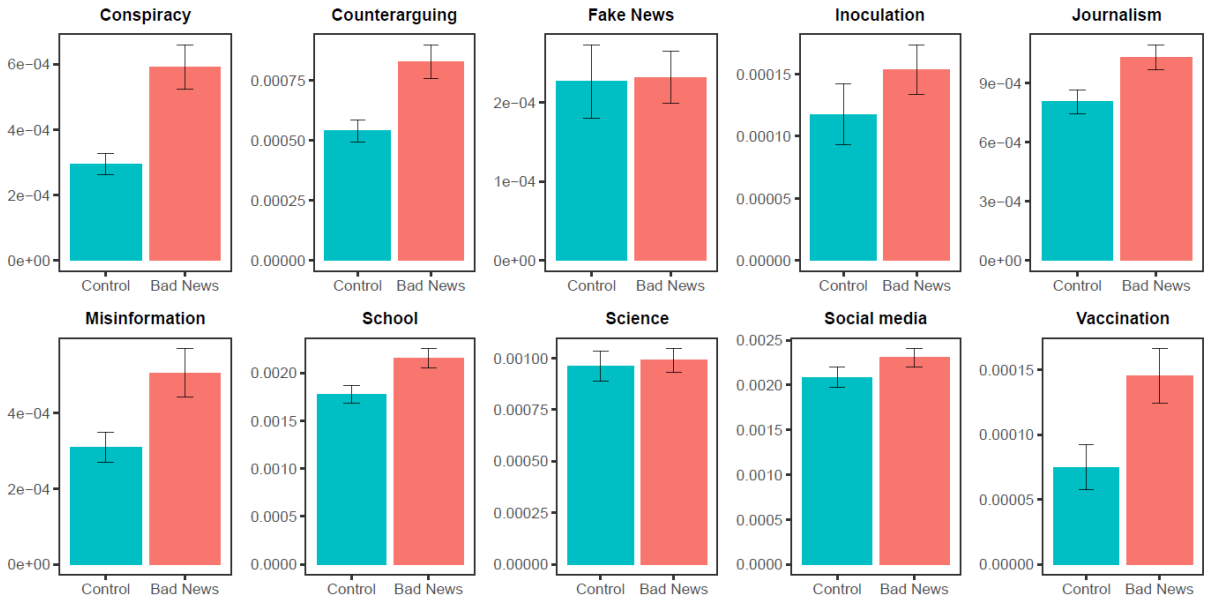

Figure 3. Bar plots, per Empath category. Error bars represent standard error. *Y*-axes show normalized Empath scores.

Category	Statistic	df	*p*	M_{diff}	95%CI	Cohen's *d*
conspiracy***	3.932	29624	< .001	2.96e-04	[0.00015, 0.00044]	0.039
counter_misinfo***	3.433	34177	< .001	2.89e-4	[1.24e-4, 4.54e-4]	0.035
fake_news	0.0857	28630	0.932	4.87e-06	[-1.06e−4, 0.00012]	9.29e-04
inoculation	1.130	31602	0.258	3.58e-05	[-2.63e−5, 0.000098]	0.012
journalism**	2.588	35655	0.01	2.25e-04	[0.000055, 0.00040]	0.027
misinformation**	2.609	33313	0.009	1.97e-04	[0.000049, 0.00035]	0.026
school**	2.776	35979	0.006	3.80e-04	[0.00011, 0.00065]	0.029
science	0.297	30775	0.766	2.78e-05	[-1.55e−4, 0.00021]	0.0032
social_media	1.440	33444	0.150	2.19e-04	[-7.92e−5, 0.00052]	0.015
vaccination**	2.601	36120	0.009	7.07e-05	[0.000017, 0.00012]	0.027

Table 2. Welch's *t*-tests for *Bad News* Empath categories. *Note*: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Figure 3 and Table 2 show that Reddit users who commented on the *Bad News* thread used language related to *conspiracy, counter_misinfo* (counterarguing against misinformation)*, journalism, misinformation, school,* and the topic of *vaccination* significantly more in the week after posting on the thread than the control group. However, we find no significant difference between the treatment and control group for the *fake news*, *inoculation, science* and *social media* categories (though the first three are statistically significant in the non-parametric Mann-Whitney U-test, see Table S1).

To explore whether time is a factor in the prevalence of issue-relevant postinoculation talk, we conduct a series of ANOVAs with each Empath category as the dependent variable and condition and time elapsed after the *Bad News* thread was posted (in days) as independent variables. We find no significant interactions between condition and the time after posting (aside from a small significant interaction for the *journalism* category, $p = 0.045$); see Table S4.

Finally, as a robustness check and to see whether the results from Table 2 are disproportionately influenced by individual Reddit users in the sample, we constructed a Linear Mixed Model with condition (treatment – control) as a fixed effect and Reddit users as a random effect and with each Empath category as a separate dependent variable. Doing so shows that the significant effects reported above remain robust when controlling for individual users (all *p*s < 0.04), except for the

vaccination category, which is near-significance ($p = 0.076$); see Table S5 for a full overview. Figure S4 shows the jitter plots for the user-averaged linear mixed models.

These results offer tentative, but not unambiguous, support for hypothesis **H1**: after inoculation, Reddit users engaged in postinoculation talk specifically about misinformation, conspiracy theories, journalism (and media), and education. Importantly, users also engaged more in counterarguing against misinformation. In addition, we find that *Bad News* commenters used significantly more language specifically about vaccination. At the same time, we do not find that *Bad News* commenters engage significantly more in talk related to social media, science or inoculation. Furthermore, the effect sizes are small, which may be a consequence of the fact that Reddit users can talk about anything at all on the platform, in any subreddit, and so the data is by definition noisy.

Discussion

Overall, we find that in an online environment where postinoculation talk is facilitated and encouraged organically, such as in Reddit threads about inoculation interventions, people eagerly discuss not only the topic of the inoculation, but also inoculation theory itself. Numerous Reddit users participating in the *Bad News* thread commented on the idea of using cognitive "vaccines" as a way to combat misinformation, with a majority appearing positively inclined towards the idea. In addition, users in both threads engaged critically with the concept of psychological inoculations, and offered their own suggestions and commentary, in what may be called "meta-inoculation talk": talk about the concept of inoculation theory itself, and how it may be applied to a pressing societal problem such as fake news. Importantly, meta-inoculation talk or talk about the importance of inoculation theory itself is a different concept from meta-inoculation, which is concerned with inoculating people against an impending inoculation treatment (Banas & Miller, 2013). This leads to important new questions, such as whether talk about inoculation theory itself could generate resistance through an understanding of its mechanisms. Consider, for example, the following quote from a user who commented on the *Bad News* thread:

> *I agree that this [game] is great as a "vaccine" to give people the toolkits of digital media literacy before they ever encounter disinformation campaigns. If you did try to create a game to get people out of entrenched conspiracies and disinformation spirals it would have to be something different. Debunking is hard and I agree that going point by point is ineffective. But I know there has been some recent work that using similar tactics as the disinformation campaigns can be effective in getting people out of those toxic frameworks (i.e., use emotion, mobilize group attitudes that create peer pressures, tell stories, use visually exciting media, etc.) I could imagine a game being a good way to do that!*

By being involved in the *Bad News* Reddit thread, this Reddit user has learned about the concept of psychological inoculations, and how they differ from the more commonly used strategy of *debunking* misinformation; for example, the user is aware that inoculation treatments are aimed at preventing unwanted persuasion rather than undoing persuasions post hoc (Compton, 2013). In addition, the *Bad News* inoculation appears to have prompted the user to think critically about the problem of misinformation, and how games may be used as a tool to mitigate it. Thus, as also exemplified by numerous examples mentioned above, we show that inoculation treatments can induce "meta-inoculation talk" on social media, in the form of critical engagement with the tenets of inoculation theory. Interestingly, the above post uses the term "vaccine" rather than "inoculation". This observation may help explain some ambiguity around the results (in that the statistical analyses revealed significantly more language use around the category "vaccination" but not consistently for "inoculation"). This may be related to the fact that in popular media, the term "fake news vaccine" has often been used, as opposed to "inoculation" (e.g., see BBC, 2017; Reuters, 2018; CNN, 2019).

In a first exploration of postinoculation talk "in the wild" on social media, we also find preliminary support that Reddit users who actively participated in an inoculation treatment indeed engage in significant postinoculation talk and counterarguing against misinformation in the week after their exposure to the treatment, when compared to a control group of Reddit users with similar interests. Although qualitative, we find that some "inoculated" users display psychological resistance against misinformation techniques learned in the *Bad News* game. For example, some treatment group participants subsequently called out other Reddit users' use of conspiratorial reasoning, which represents a major misinformation technique that players learn about in the game (Roozenbeek & van der Linden, 2019): one user counterargued against other users' supposed conspiratorial reasoning in a subreddit about Game of Thrones, a few days after posting in the *Bad News* thread: "*Yes yes, it's all some giant conspiracy involving Disney. You people who take this stuff seriously are hilarious.*" Another user appears to have been motivated to debunk another Reddit user's comment: "*I love how you're extrapolating this nonsense from a post that has no evidence of that.*" Such individual comments are qualitative, however, and hardly indicative of broader patterns; we have provided some initial evidence that patterns of increased issue-relevant postinoculation talk are present among *Bad News* commenters when compared to a control group.

Having said this, our data is not granular enough to examine more specific postinoculation talk for each of the misinformation techniques featured in the *Bad News* game, and we urge a cautious interpretation of our findings. We therefore highly encourage further research into how (and if) postinoculation talk occurs "in the wild" on social media, and how inoculation messages proliferate through social media platforms.

Of course, our study is not without limitations. Because we could not link Reddit usernames to survey data due to General Data Protection (GDPR) regulations, our study design did not allow us to analyze whether postinoculation talk bolstered people's resistance to future persuasion attacks (Ivanov *et al.*, 2012), or strengthened belief certainty (Dillingham and Ivanov, 2016) about the target issue; using "big data" linguistic analysis tools such as Empath have a downside in that they are unable to clarify in what context certain language was used. For a more detailed discussion about this limitation, we refer to Fong et al. (2021). We were also unable to check if Reddit users shared less misinformation on the platform than a control group, for a variety of reasons: first, Reddit communities are usually heavily moderated, and misinformation posts often get deleted, particularly about controversial topics (Mak, 2020). Second, although the number of posts was sizable ($N > 1{,}000$) and suitable for linguistic analyses, the number of users was not enough to be able to detect meaningful differences between groups to determine whether inoculated Reddit users spread or interacted with less misinformation than the control group on the platform. Further research is needed to explore these questions in depth.

Furthermore, although we find significant effects for postinoculation talk, the effect sizes are very small (between $d = 0.026$ and $d = 0.037$). However, it may simply be too optimistic to expect large effects, as social media field experiments often report small between-group differences (Pennycook *et al.*, 2021); over thousands of posts, the impact of individual posts in the dataset may be quite low, even if the effects are significant. Klein et al. (2019), for example, use a threshold of $d = 0.20$ as a cut-off point for effect size for between-group Empath category comparisons, which is also in the low range for psychological research (Funder and Ozer, 2019); it is important to note that Reddit research often compares different subreddits to each other (Roozenbeek and Salvador Palau, 2017), and not groups of individual users (as we did), which in our case may further reduce the effect sizes. In addition, and as exemplified in Figure S3, some Empath categories are simply not common in our corpus of Reddit posts, which leads to low base values and hence low mean normalized Empath scores, regardless of experimental condition. Thus, it may be the case that low effect sizes are to be expected when using linguistic dictionaries (Fong *et al.*, 2021).

Finally, once the *VaxBN* and *Bad Vaxx* games are launched in the public domain there might be future opportunities for examining more vaccine-specific inoculation talk. Nonetheless, despite these limitations, we find that studying how social media users engage not only with inoculation treatments but with inoculation *theory* more generally (as well as increased counterarguing against the topic of the inoculation, in our case misinformation) can contribute to our understanding of how inoculations can spread through interpersonal communication.

Conclusion

Although mostly exploratory, we emphasize the value of studying postinoculation talk "in the wild" on social media, and we have made several methodological and analytical strides towards the study of this important topic. Active involvement with a particular topic, either through gaming or by engaging in an interactive discussion, prompts independent further communication about the topic of discussion. Specifically, we show that active engagement with inoculation treatments may indeed induce both postinoculation talk (Compton and Pfau, 2009), issue-relevant counterarguing, and what we here call "meta-inoculation talk" outside of a laboratory setting. We thus offer a way forward for researchers to look at postinoculation talk in social media environments in other issue domains where inoculation treatments have proven their effectiveness and where a great deal of debate takes place on social media, including debates around climate change (van der Linden, Leiserowitz, *et al.*, 2017), risky health behaviors (Parker, Ivanov and Compton, 2012) and conspiracy theories (Banas and Miller, 2013).

Most importantly, this study shows that inoculation interventions such as *VaxBN* and *Bad Vaxx* (developed as part of WP2.2) have the potential to generate significant independent discussion on social media, not only about the interventions themselves but about inoculation theory in general and specifically within the context of misinformation. This finding is important because it demonstrates that discussions about the intervention and the issues that they seek to tackle (in this case vaccine misinformation) do not stop when the intervention is completed. Rather, people may continue to independently discuss the lessons they learned in the games with other people that they encounter online. Although tentative, this offers exciting opportunities for the feasibility of so-called "psychological herd immunity" against misinformation (Pilditch *et al.*, 2022).

Next steps

Future research may explore how inoculated individuals' behavior on social media differs from non-inoculated individuals, in terms of both sharing misinformation and the inocula tion; in addition, future work may tackle postinoculation talk specifically about vaccine misinformation and vaccine-specific inoculation interventions. These questions will be examined as part of WP3.2, where JITSUVAX researchers will look at postinoculation talk on a simulated social media platform, Mastodon. Doing so will bring about much-needed insights into the effectiveness of psychological inoculation in controlled environments, and offers a way forward for exploring relatively understudied concepts within inoculation theory such as the potential for pass-along effects and societal "herd immunity" (Compton and Pfau, 2009; van der Linden, Maibach, *et al.*, 2017; Rains, 2018). These insights are highly relevant not only for the field of communication research, but also for the scalability of inoculation interventions, and the feasibility of a "broad-spectrum" vaccine against misinformation.

References

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

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

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

Basol, M. *et al.* (2021) 'Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation', *Big Data and Society*, 8(1). doi: 10.1177/20539517211013868.

Basol, M., Roozenbeek, J. and 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), pp. 1–9. doi: https://doi.org/10.5334/joc.91.

BBC (2017) *Cambridge scientists consider fake news 'vaccine'*, *BBC News*. Available at: http://www.bbc.co.uk/news/uk-38714404 (Accessed: 29 August 2017).

BBC News (2018) 'A fake billionaire is fooling people on Twitter', *www.bbc.co.uk*, 28 August. Available at: https://www.bbc.co.uk/news/world-us-canada-45331781 (Accessed: 12 December 2018).

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

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) 'Latent Dirichlet Allocation', *J. Mach. Learn. Res.* JMLR.org, 3, pp. 993–1022. Available at: http://dl.acm.org/citation.cfm?id=944919.944937.

Brady, W. J. *et al.* (2017) 'Emotion shapes the diffusion of moralized content in social networks', *Proceedings of the National Academy of Sciences*, 114(28), pp. 7313–7318. doi: 10.1073/pnas.1618923114.

CNN (2019) 'Researchers have created a "vaccine" for fake news. It's a game', *edition.cnn.com*, 4 July. Available at: https://edition.cnn.com/2019/07/04/media/fake-news-game-vaccine/index.html.

Compton, J. (2013) 'Inoculation Theory', in Dillard, J. P. and Shen, L. (eds) *The SAGE Handbook of Persuasion: Developments in Theory and Practice*. 2nd edn. Thousand Oaks: SAGE Publications, Inc., pp. 220–236. doi: 10.4135/9781452218410.

Compton, J. (2019) 'Prophylactic versus therapeutic inoculation treatments for resistance to influence', *Communication Theory*, 30(3), pp. 330–343. doi: 10.1093/ct/qtz004.

Compton, J. and Ivanov, B. (2013) 'Vaccinating voters: Surveying political campaign inoculation scholarship', *Annals of the International Communication Association*, 37(1), pp. 251–283. doi: 10.1080/23808985.2013.11679152.

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

Compton, J. and Pfau, M. (2005) 'Inoculation Theory of Resistance to Influence at Maturity: Recent Progress In Theory Development and Application and Suggestions for Future Research', *Annals of the International Communication Association*. Routledge, 29(1), pp. 97–145. doi: 10.1207/s15567419cy2901_4.

Compton, J. and Pfau, M. (2009) 'Spreading Inoculation: Inoculation, Resistance to Influence, and Word-of-Mouth Communication', *Communication Theory*, 19(1), pp. 9–28. doi: 10.1111/j.1468-2885.2008.01330.x.

Cook, J., Lewandowsky, S. and Ecker, U. K. H. (2017) 'Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence', *PLOS ONE*. Public Library of Science, 12(5), pp. 1–21. doi: 10.1371/journal.pone.0175799.

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

Curiskis, S. A. *et al.* (2019) 'An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit', *Information Processing & Management*. doi: https://doi.org/10.1016/j.ipm.2019.04.002.

Dillingham, L. L. and Ivanov, B. (2016) 'Using Postinoculation Talk to Strengthen Generated Resistance', *Communication Research Reports*, 33(4), pp. 295–302.

Eagly, A. H. and Chaiken, S. (1993) *The Psychology of Attitudes*. Orlando, FL: Harcourt Brace Jovanovich.

Fast, E., Chen, B. and Bernstein, M. S. (2016) 'Empath: Understanding Topic Signals in Large-Scale Text', in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4647–4657. doi: 10.1145/2858036.2858535.

Fong, A. *et al.* (2021) 'The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter.', *Group Processes & Intergroup Relations*. doi: 10.1177/1368430220987596.

Funder, D. C. and Ozer, D. J. (2019) 'Evaluating Effect Size in Psychological Research: Sense and Nonsense', *Advances in Methods and Practices in Psychological Science*. SAGE Publications Inc, 2(2), pp. 156–168. doi: 10.1177/2515245919847202.

Gerlach, M., Peixoto, T. P. and Altmann, E. G. (2018) 'A network approach to topic models', *Science Advances*, 4(7). doi: 10.1126/sciadv.aaq1360.

Goga, O., Venkatadri, G. and Gummadi, K. P. (2015) 'The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks', in *Proceedings of the 2015 Internet Measurement Conference*. New York, NY, USA: ACM (IMC '15), pp. 141–153. doi: 10.1145/2815675.2815699.

Griffiths, M. D. (2014) 'Adolescent trolling in online environments: a brief overview', *Education and Health*. Exeter: Schools Health Education Unit (SHEU), 32(3), pp. 85–87. Available at: http://irep.ntu.ac.uk/id/eprint/25950/.

Isoaho, K., Gritsenko, D. and Mäkelä, E. (2019) 'Topic Modeling and Text Analysis for Qualitative Policy Research', *Policy Studies Journal*. John Wiley & Sons, Ltd. doi: 10.1111/psj.12343.

Ivanov, B. *et al.* (2012) 'Effects of Postinoculation Talk on Resistance to Influence', *Journal of Communication*, 62(4), pp. 701–718.

Ivanov, B. *et al.* (2015) 'The General Content of Postinoculation Talk: Recalled Issue-Specific Conversations Following Inoculation Treatments', *Western Journal of Communication*. Routledge, 79(2), pp. 218–238. doi: 10.1080/10570314.2014.943423.

Ivanov, B. (2017) 'Inoculation Theory Applied in Health and Risk Messaging', *Oxford Research Encyclopedia of Communication*, 24 May. Available at: https://oxfordre.com/communication/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-254.

Ivanov, B. *et al.* (2017) '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*, pp. 1–11. doi: 10.1111/1468-5973.12213.

Iyengar, S. and Massey, D. S. (2018) 'Scientific communication in a post-truth society', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 116(16), pp. 7656–7661. doi: 10.1073/PNAS.1805868115.

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

Kapadia, S. (2019) *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*, *Towards Data Science*. Available at: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 (Accessed: 11 June 2020).

Klein, C., Clutton, P. and Dunn, A. G. (2019) 'Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum', *PLOS ONE*, 14(11), p. e0225098. doi: 10.1371/journal.pone.0225098.

Klein, C., Clutton, P. and Polito, V. (2018) 'Topic Modeling Reveals Distinct Interests within an Online Conspiracy Forum', *Frontiers in psychology*. Frontiers Media S.A., 9, p. 189. doi: 10.3389/fpsyg.2018.00189.

Lewandowsky, S., Oberauer, K. and Gignac, G. E. (2013) 'NASA Faked the Moon Landing—Therefore, (Climate) Science Is a Hoax: An Anatomy of the Motivated Rejection of Science', *Psychological Science*, 24(5), pp. 622–633. doi: 10.1177/0956797612457686.

van der Linden, S. (2013) 'Why people believe in conspiracy theories (What a Hoax)', *Scientific American Mind*, 24, pp. 41–43.

van der Linden, S., Maibach, E., *et al.* (2017) 'Inoculating against misinformation', *Science*, 358(6367), pp. 1141–1142. doi: 10.1126/science.aar4533.

van der Linden, S., Leiserowitz, A., *et al.* (2017) 'Inoculating the Public against Misinformation about Climate Change', *Global Challenges*, 1(2), p. 1600008. doi: 10.1002/gch2.201600008.

van der Linden, S. and Roozenbeek, J. (2020) 'Psychological inoculation against fake news', in Greifenader, R. et al. (eds) *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*. London: Psychology Press. doi: 10.4324/9780429295379-11.

Lischka, J. A. (2017) 'A Badge of Honor?: How The New York Times discredits President Trump's fake news accusations', *Journalism Studies*, pp. 1–18. doi: 10.1080/1461670X.2017.1375385.

Maertens, R. *et al.* (2021) 'Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments', *Journal of Experimental Psychology: Applied*, 27(1), pp. 1–16. doi: 10.1037/xap0000315.

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

Mak, A. (2020) *Coronavirus Diaries: I Spend Hours Removing Outbreak Conspiracies From Reddit*, *Slate*. Available at: https://slate.com/technology/2020/03/coronavirus-diaries-reddit-moderator-conspiracy-theories.html (Accessed: 11 June 2020).

McCallum, A. K. (2002) *MALLET: A Machine Learning for Language Toolkit*, *Mallet UMass*. Available at: http://mallet.cs.umass.edu/ (Accessed: 8 June 2020).

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

McGuire, W. J. and 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, pp. 326–332.

McGuire, W. J. and Papageorgis, D. (1961b) 'The relative efficacy of various types of prior belief-defense in producing immunity against persuasion.', *Journal of abnormal and social psychology*, 62(2), pp. 327–337.

Navarro-Colorado, B. (2018) 'On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry', *Frontiers in Digital Humanities*, 5, p. 15. doi: 10.3389/fdigh.2018.00015.

Papageorgis, D. and McGuire, W. J. (1961) 'The generality of immunity to persuasion produced by pre-exposure to weakened counterarguments', *Journal of abnormal and social psychology*, 62, pp. 475–481.

Parker, K. A., Ivanov, B. and Compton, J. (2012) 'Inoculation's efficacy with young adults' risky behaviors: Can inoculation confer cross-protection over related but untreated issues?', *Health Communication*, 27(3), pp. 223–233. doi: 10.1080/10410236.2011.575541.

Pennebaker, J. W. *et al.* (2015) *Linguistic Inquiry and Word Count: LIWC 2015*. Austin, TX: Pennebaker Conglomerates.

Pennycook, G. *et al.* (2021) 'Shifting attention to accuracy can reduce misinformation online', *Nature*, 592, pp. 590–595. doi: s41586-021-03344-2.

Pilditch, T., Roozenbeek, J., Madsen, J., & van der Linden, S. (2022). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science, 9*(211953). https://doi.org/10.1098/rsos.211953

Prabhakaran, S. (2018) *Topic Modeling with Gensim (Python)*, *Machine Learning Plus*. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/ (Accessed: 8 June 2020).

Prabhakaran, S. (2019) *Topic modeling visualization - How to present the results of LDA models?*, *Machine Learning Plus*. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/ (Accessed: 8 June 2020).

Rains, S. A. (2018) 'Big data, computational social science, and health communication: A review and agenda for advancing theory', *Health Communication*, pp. 1–9. doi: 10.1080/10410236.2018.1536955.

Reddit (2019a) *Fake news 'vaccine' works, suggests a large new study (n=15,000), which shows a simple online game works like a "vaccine", increasing skepticism of fake news by giving people a "weak dose" of the methods behind disinformation, a version of what psychologi*, *www.reddit.com*. Available at: https://www.reddit.com/r/science/comments/c5ptfz/fake_news_vaccine_works_suggests_a_large_new/ (Accessed: 16 May 2020).

Reddit (2019b) *People who spend at least 2 hours in nature a week are significantly more likely to report good health and higher psychological wellbeing, according to a new large-scale study (n = 19,806), which found that it didn't matter whether this was achieved in a* , *www.reddit.com*. Available at: https://www.reddit.com/r/science/comments/c0igz7/people_who_spend_at_least_2_hours_in_nature_a/ (Accessed: 5 June 2020).

Reuters (2018) *Fake news 'vaccine' teaches you to spot disinformation*, *www.uk.reuters.com*. Available at: https://uk.reuters.com/video/2018/03/20/fake-news-vaccine-teaches-you-to-spot-di?videoId=410596269 (Accessed: 15 January 2019).

Reznik, M. (2013) 'Identity Theft on Social Networking Sites: Developing Issues of Internet Impersonation', *Touro Law Review*, pp. 455–484.

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

Rojecki, A. and Meraz, S. (2016) 'Rumors and factitious informational blends: The role of the web in speculative politics', *New Media & Society*, 18(1), pp. 25–43. doi: 10.1177/1461444814535724.

Roozenbeek, J. (2020) 'Identity discourse in local newspapers before, during and after military conflict: a case study of Kramatorsk', *Demokratizatsiya: The Journal of Post-Soviet Democratization*, 28(3), pp. 419–459. Available at: https://muse.jhu.edu/article/762316/pdf.

Roozenbeek, J. *et al.* (2021) 'Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation.', *Educational and Psychological Measurement*, 81(2), pp. 340–362. doi: 10.1177/0013164420940378.

Roozenbeek, J. and van der Linden, S. (2018) 'The fake news game: actively inoculating against the risk of misinformation', *Journal of Risk Research*, 22(5), pp. 570–580. doi: 10.1080/13669877.2018.1443491.

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

Roozenbeek, J., van der Linden, S. and Nygren, T. (2020) 'Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures', *The Harvard Kennedy School (HKS) Misinformation Review*, 1(2). doi: 10.37016//mr-2020-008.

Roozenbeek, J. and Salvador Palau, A. (2017) 'I read it on reddit: Exploring the role of online communities in the 2016 US elections news cycle', in *Proceedings of the 2017 International Conference on Social Informatics*, pp. 192–220. doi: 10.1007/978-3-319-67256-4_16.

Saleh, N. *et al.* (2021) 'Active inoculation boosts attitudinal resistance against extremist persuasion techniques – A novel approach towards the prevention of violent extremism', *Behavioural Public Policy*, pp. 1–24. doi: 10.1017/bpp.2020.60.

Walton, D. (1998) *Ad Hominem Arguments*. Tuscaloosa and London: The University of Alabama Press.

White, M. P. *et al.* (2019) 'Spending at least 120 minutes a week in nature is associated with good health and wellbeing', *Scientific Reports*, 9(1), p. 7730. doi: 10.1038/s41598-019-44097-3.