After the Shock
Reform, Resilience, and Economic Transformation in MENA
ERF 32nd Annual Conference
June 14-16 | Cairo, Egypt

2026

Can Individuals Be Inoculated against Fake News:

An Experimental Study on a Sample of Egyptian X Users

Mazen Hassan,
Sarah Mansour,
Asmaa Abd El Khalek,
Mohamed Sabry Amer
and Zeyad Kelani

ECONOMIC RESEARCH FORUM
منتدى البحوث الاقتصادية

Can Individuals Be Inoculated against Fake News:
An Experimental Study on a Sample of Egyptian *X* Users

Abstract

The spread of fake news has been shown to have alarming political and social consequences. In this paper, we examine whether providing a short training to social media users can help them resist fake news, measured by how far they believe them and their tendency to share them. Our design tests a hypothesis which assumes that individuals can be inoculated with mental antibodies against fake news – just as individuals get vaccinated against viruses. We test this hypothesis via a pre-registered online experiment, conducted on Egyptian *X* users (formerly Twitter). We recruited our subjects from a sample of trend engagers who interacted with the top trending keywords/hashtags over two months, in addition to a booster sample of university students. The experiment entailed randomly assigning subjects into a control group, a treatment that trained them on scientific reasoning, and another that briefed them on the possibility of inflicting social harm when sharing fake news. The findings show that both treatments were effective in reducing the belief in fake news, whereas only the scientific treatment was effective in curbing the tendency to share such news. The paper contributes to the literature on resisting misinformation by testing the inoculation theory on actual X users; a traditionally under-researched sample.

I. Introduction.

Over the past few years, the likelihood of exposure to fake news – defined as content that resembles legitimate news but is however fabricated or extremely inaccurate (Pennycook and Rand 2021)[1] – has been rising significantly (Newman et al. 2021). Contributing factors include the increased usage of social media worldwide[2], the growing reliance on digital outlets as sources of news, and the enhanced technological capacity to produce falsified news with the ascend of artificial intelligence (Simon et al 2023). Increased exposure, in turn, means that traces of fake news are likely to find their way into the large pool of information individuals rely on when making crucial decisions. Indeed, fake news have been shown to influence voting decisions (Bennett and Livingstone 2018), attitudes towards government (Linden, Leiserowitz, Rosenthal, and Maibach 2017), stock prices, vaccination rates (Larson et al 2011), and sometimes were directly responsible for loss of lives (Arun 2019).

Within this context, this paper examines how to increase resistance to fake news among social media users, being a group that is probably more likely to be exposed to fake news. Specifically, we test whether a preemptive immunity – triggered by experimental interventions – could be generated to increase individuals' resistance to fake news (Compton 2013; Banas and Rains 2010). We measure such resistance by two variables: believability of fake news and tendency to share it. Our hypothesis is that generating a preemptive immunity is possible by giving *prior doses* of fake news to individuals, either via training them on scientific reasoning (our first treatment/hypothesis) or highlighting the potential of social harm that might result from sharing fake news (second treatment).

[1] Other detailed definitions distinguish between related sub-terms as disinformation (where the intention to produce false news is the defining factor), misinformation (where such intention is absent), and falsified news (where the intention to *harm* is the defining factor).

[2] https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

We focus on social media (particularly *X*) users who interacted with top hashtags/keywords in Egypt (*X* users make up 41% of our subjects, among whom 57% are Egyptians whereas the rest from other Arab countries).[3] Our case selection has multiple reasons. *First*, most of the literature on fake news and misinformation tends to focus on democratic (and mainly Western industrialized) contexts. This is largely unjustified given that fake news are likely to spread faster in authoritarian countries due to lower levels of trust in government and less government transparency (Bateman and Jackson 2024). Our focus on an Arab-subject pool therefore helps us make contributions to the literature by testing our hypotheses in a region that is traditionally, and persistently, authoritarian. *Second*, the Arab World – and Egypt as the most populous Arab country – is also a region with high illiteracy rates,[4] low levels of political engagement (which is usually a natural trainer of individuals on filtering news), and a more-than-average spread of belief in conspiracy theory (Zonis and Joseph 1994). The combination of these factors makes our case a fertile milieu for the spread of fake news (van Prooijen 2017; Lavigne et al 2022).

Third, although with notable exceptions (Roozenbeek et al 2022), many misinformation studies tend *not* to test their hypotheses on active social media users because of the complexity of running experiments on such segment. Social media users on the other hand are likely to be most exposed to fake news (Lazer at al 2018). Our third contribution therefore is that we recruit almost half of our subjects (41%) from amongst *X* users who engage with the trending keywords and hashtags. This design feature is likely to increase our study's ecological validity and empirical contribution. *Fourth*, Egypt is a country where multiple interventions against fake news have been

[3] We also targeted *X* users in Saudi Arabia as part of our sample. The aim was to include Egypt and Saudi Arabia in the study, especially given that Saudi Arabia is the Arab country with the highest percentage of its population active on X. However, the responses we got from *X* users who interacted with trending Saudi hashtags/keywords (our main tool to target potential subjects as will be shown later) was too low, as shown in the analysis section.

[4] https://observatory.alecso.org/Data/wp-content/uploads/2023/05/nashra9.pdf

tested (including a law and a government observatory unit issuing daily corrections), [5] but where there is still public (and official) acknowledgment and incidents indicating a high level of fake news circulation[6] – making Egypt a suitable case to test hypotheses on prior immunity.

To test our hypothesis, we conducted an online experiment on a sample of *trend engagers* on *X* (*X* users who interacted with trending keywords and hashtags over two months) and an additional booster sample of students at an Egyptian public university. We started our recruitment process by identifying top trending keywords and hashtags in Egypt on *X*, and then used *X*'s API to identify users who interacted with these keywords and hashtags. In total, we managed to scrape 21,697 *X* users using this method. After filtering irrelevant accounts (those using foreign languages, suspected bots, and institutional accounts), we invited our subjects to an online survey. As we expected a low response rate of less than 1%, we simultaneously ran the experiment with a booster sample of university students. In total we managed to recruit 288 respondents (119 *X* users and 169 students). Our reported results control for whether subjects are X users or students.

The experimental design randomly assigned subjects to a control group, or one of two treatment groups. Subjects in the control group were asked to read a news report that contained neutral information, whereas in the first treatment, subjects were presented with a text showing how scientific reasoning can disprove prior widely circulated explanations of social phenomena (we call this treatment *scientific reasoning*). In the second treatment, subjects were given a text showing how violence could erupt due to the circulation of false news (*social harm* treatment). After exposure to such texts, individuals were asked to determine their believability of, and

[5] https://www.almasryalyoum.com/news/details/3342022#google_vignette

[6] https://www.youm7.com/story/2025/2/16/%D8%A7%D9%84%D8%AD%D9%83%D9%88%D9%85%D8%A9-%D8%A7%D8%B1%D8%AA%D9%81%D8%A7%D8%B9-%D9%86%D8%B3%D8%A8%D8%A9-%D8%A7%D9%84%D8%B4%D8%A7%D8%A6%D8%B9%D8%A7%D8%AA-%D9%84%D9%80-16-2-%D8%B9%D8%A7%D9%85-2024-%D9%85%D9%82%D8%A7%D8%B1%D9%86%D8%A9%D9%8B/6885372

tendency to share, four pieces of news that were created by the authors and tested on a pre-sample (three of them were fake news stories and one was correct).

The results show that both treatments significantly reduced subjects' average belief in fake news (but did not affect their rating of the correct news). When it comes to the tendency to share, the *scientific reasoning* also produced significant effects whereas the *social harm* treatment lacked significance (although generated movement in the expected direction). These findings contribute to the literature linking prior training and inoculation theory with fake news resistance (Hameleers 2022; Compton 2013; Banas and Rains 2010), as well as the strand of the literature trying to test fake news interventions on actual social media users (Roozenbeek et al 2022), rather than via lab experiments. By focusing on subjects mainly coming from authoritarian contexts, we also test our hypothesis on an under-researched sample and one that is more likely to get exposed to fake news. Our significant results are also important given the difficulty of reaching statistically significant results in empirical studies on fake news (IPIE 2023; Chan et al 2017).

From this point, this paper proceeds as follows. The next section presents our theoretical argument whereas section three presents the methodology. Section four presents the results and section five concludes.

II. Social Inoculation Theory.

According to the misinformation literature, four approaches are usually suggested to reduce the circulation of fake news. The first is predominated by a sphere of social media companies where *algorithms* are developed to detect and label dubious news on social media platforms, as a way of reducing the possibility of sharing them by users (Clayton et al 2019). The second approach relies on ex-post *fake news corrections* where media outlets – or government agencies – constantly

check circulated news and issue warnings highlighting potentially false ones (Nyhan 2019). The third tool is *legislative* according to which laws are drafted to criminalize generating and spreading fake news. Finally, the fourth approach rests on *psychological theories*, and envisions – as one of its techniques – training individuals to sort news into potentially fake ones or of reasonable credibility (Banas and Rains 2010).

In this paper, we focus on the fourth *psychological* approach for the following reasons. Firstly, whereas the algorithmic labelling, ex-post correction and the legislative approaches seem to fall largely into the sphere of social media companies and/or governments, multiple studies do suggest that correction attempts significantly lose their effectiveness after short time periods (Walter and Murphy 2018; Cook, Ecker, and Lewandowsky 2015; Lewandowsky, Ecker, Seifert, Schwarz, and Cook 2012; Schwarz, Newman, and Leach 2016), or sometimes even backfire by increasing belief in fake news (Nyhan and Reifler 2012). Developing a preemptive fake news immunity, on the other hand, would enable individuals to resist fake news independent of ex-ante interventions by social media companies or ex-post corrections by governments (which is highly relevant in authoritarian countries where government could themselves be the source of fake news). Moreover, especially in the Arab context – the focus of this paper – ex-post corrections are likely to be even more challenging because of the spread of a 'culture of honor' (Nisbett 1996) that makes individuals dislike to be proven wrong, or have their beliefs checked and corrected by others. In such contexts, individuals may intentionally avoid corrective interventions contradicting with their prior beliefs (Hameleers and Van der Meer 2019) or disapprove of them.

Our argument therefore is one that examines the notion of potential preemptive resistance: that generating a fake news immunity – via social inoculation – would create *mental antibodies* in the same way that immunization protects individuals against viruses (Compton 2013; Banas and

Rains 2010; Maertens et al. 2021). Whereas social inoculation theory was first coined by McGuire (1964) during the height of the Cold War to investigate how best to enhance individuals' abilities to resist disinformation, it has also attracted recent attention (van der Linden and Roozenbeek 2021). According to the multiple versions of this theory, for inoculation to work, it needs to have two components: *persuasion* and *attitude change*. We designed two treatments to test each of these components.

The mental, persuasion component of inoculation theory aims to equip individuals with tools of scientific reasoning, as a way to enable them to challenge potentially fake news. The idea here is that most fake news suffers from reasoning problems: that the primary message or conclusion is not necessarily a logical result of the facts given to make up the story (Bowers 2021). Training individuals to check and challenge the link between the cause and effect therefore is one way to detect lies and potentially fake content (Lewandowsky and van der Linden 2012; Roozenbeek and van der Linden 2019).

Several studies do support the importance of sound critical thinking when absorbing news (see van Prooijen 2017). Livingstone also pointed out that prior training on the ability to spot fake photos and videos, as well as applying an "if…then" mentality, increases audience engagement in fact-checking processes and reduces the spread of misinformation (Livingstone 2022; see also Armeen, Niswanger, and Tian 2024). In the Arab context, such intervention is highly relevant given what several studies show with regard to embeddedness of groupthink and indoctrination techniques in many Arab educational systems (Salih 2009; Al-Hadabi and Al-Shawal 2012; Al-Mahrooqi and Denman 2020). Accordingly, we designed our first treatment (*scientific reasoning*) in a way that trains subjects to critically assess information – as opposed to accepting content at

face value – by showing them how the first and fast conclusions drawn from data could be wrong. Our first hypothesis thus is as follows:

H_1: *Training individuals to use scientific reasoning when consuming news will reduce their tendency to believe - and share - fake news.*

The second component projected by inoculation theory is *psychological*, related to *attitude change*. Its main assumption is that triggering emotions (especially negative ones, like threat, harm, etc.) is a strong driver of human action, including sharing and circulating fake news. We make use of this component in our second treatment by highlighting the potential of harm that could be inflicted on others when circulating fake news, as a means to contain its spread. Our argument is that emphasizing the possibility of causing social harm reverses the effect of one commonly known driver for sharing fake news on social media in the first place: the desire for social interactions.[7] Indeed, several studies have indicated that individuals tend to share fake news as a means of attracting attention in the digital sphere. In today's world, where social media is the largest space to meet others – and for some, perhaps the primary one for forming social relationships – individuals learn quickly that an effective way to gain attention is by sharing news of non-standard content (Sung et al. 2016; Burrow and Rainone 2017; Steers et al. 2016). A recent study that focused on users' motivational structures (Globig, Holtz, and Sharot 2023) did show that the desire for virtual social acceptance is linked to the news being circulated.

Our second intervention therefore taps into such driver by highlighting that sharing fake news could cause harm or damage to others and potentially one's network of relationships, hence

[7] There are certainly several other motivations that drive individuals to circulate and share news on social media platforms (see Berge and Milkman 2012), including, for example, the desire to spread one's ideas and the bias in favour of negative news (Soroka 2014; Haas et al. 2020).

contradicting the social purpose of sharing news. The underlying logic behind the intervention is that a person – motivated by a desire to form social relationships – would make a simple calculus if trained on potential harm of sharing news of questionable credibility, that would ultimately lead them to reduce the sharing of such news. Our second hypothesis thus reads as follows:

H_2: *Highlighting the social harm that may result from circulating and sharing fake news will reduce individuals' tendency to believe - and to share - fake news.*

III. Methodology – Identifying Trend engagers on *X* and Experimental Design.

We conducted a pre-registered online survey experiment to test our two hypotheses.[8] In this section, we explain (a) how subjects were identified and recruited, and (b) the experimental design.

A. Identifying and contacting *trend engagers* on *X*.

To increase our study's ecological validity, we recruited our subjects from amongst social media users on *X* who interacted with top keywords and hashtags over a period of two months – whom we thus call *trend engagers*. We aimed to recruit subjects from such a population because they would be presumably frequently exposed to fake news. They are therefore the kind of subjects that would make our experiment as close to real life as possible (Morton and Williams 2010). Moreover, very few previous studies have relied on such category of respondents, given the difficulties associated with targeting them (as will be mentioned later in detail). Recruiting almost half of our subjects from such a pool hence increases the contribution this paper seeks to make.

[8] Pre-registration process included information on the main research question, key hypotheses of the study, a description of dependent variables and how they will be measured, the number of conditions subjects will be assigned to and the type of analyses to be conducted. An anonymized copy of the pre-registration, created by the authors to use during peer-review, can be found in the supplementary material.

We chose *X* as our social media platform because of three reasons. On the one hand, *X* – along with *Facebook* – are the two most used platforms among Arab republics (Reyaee and Ahmed 2015). On the other hand, multiple reports indicate that *X* does seem to have a relatively high rate of fake news circulation compared to other social media platforms, making it suitable for a study testing interventions to increase resistance to fake news.[9] Moreover, X – unlike *Facebook* – allows access to its application programming interface (API), enabling the extraction of data about interactions on the platform, which is necessary to identify trend engagers.

Previous studies have relied on different approaches to identify users of interest on *X*. While some studies focused on locating such users by focusing on specific themes (e.g. sports, see Şimşek and Kabakuş 2018), others chose to build their own social networks, and identify the most prominent users on these private networks (Alp and Öğüdücü 2018; Nebot et al. 2018) – a design that certainly reduces the ecological validity of research outcomes. In this paper, we identified trend engagers on *X* based on specific topics and issues, while relying on a diverse set of topics. To identify our subjects and contact them, we followed a three-step methodology, where in the first step we identified the list of trending keywords and hashtags on *X* and then extracted data on users who engaged with such keywords and hashtags via *X*'s API. In the third step, we created accounts on *X* to contact identified *X* users. These three steps are explained below.

Step 1 – Selection of Keywords and Hashtags. We started by identifying the top 20 keywords and hashtags[10] that Egyptian *X* users interacted with over around two months (from February 15th to April 28th, 2024). This exercise produced 1,307 keywords and hashtags on

[9] https://www.theguardian.com/technology/2023/sep/26/eu-warns-elon-musk-that-twitter-x-must-comply-with-fake-news-laws#:~:text=The%20EU%20has%20issued%20a,all%20large%20social%20media%20platforms

[10] *X* calculates engagement rates by considering two types of variables. The first type is keywords, which are words that appear repeatedly in the posts circulating on the platform and are not preceded by a hashtag symbol #. The second type of variables is the hashtags or tags, which are words or texts preceded by the hashtag symbol #.

different topics that we then classified into four groups: socio-economic, political, sports, and entertainment topics (see table 1). The distribution of keywords and hashtags over these four topics does show a reasonable degree of diversity among topics. Sports topics appear to be overrepresented, especially compared to political ones, mostly because sports topics tend to dominate user interactions on Arabic *X*. Additionally, as the X API search engine is wired to search for tweets under certain hashtags and keywords for an entire week, we eliminated similar trending hashtags in the same week. For example, if "#the_weekend" was trending on Friday, Saturday, and Sunday of the same week, we added it to our list of trending hashtags on Friday and excluded it from our list on Saturday and Sunday.

Table 1 - Topic distribution of extracted keywords and hashtags

Topic	No. of Keywords and Hashtags
Social and Economic Topics	200
Political Topics	171
Sports Topics	507
Entertainment Topics	429
Total	**1,307**

Step 2 - Extracting data of trend engagers via X API. The second step was to identify trend engagers who interacted with the keywords and hashtags identified in the previous step. This was done using *X's* API – which is an interface offered by *X* to facilitate the extraction of data across the platform. Since the API data do not provide the geographic location of the tweet creator, we relied on the language of the data extracted (i.e. Arabic) as a proxy for the Arab identity of the user. Acknowledging that this certainly cannot be the only criterion, we thus asked respondents in the post experiment questionnaire about nationality (as will be shown later).[11] At the end of this

[11] Our sample of trend engagers came from the following countries: Egypt (57%), Saudi Arabia (13%), Yemen (9%), (4%), Jordan (3%), Sudan (3%), Lebanon (3%), Kuwait (3), Morocco (2%) and Libya (2%).

stage, a total of 26,987 tweets were scraped, belonging to 21,697 unique trend engagers. This is because several trend engagers were scraped multiple times for different tweets. The distribution of trend engagers across the four topics is shown in table 2 below.

Table 2 - Distribution of trend engagers by topics of hashtags/keywords

Topic	No. of users
Social and Economic Topics	2,905
Political Topics	3,588
Sports Topics	8,618
Entertainment Topics	7,455
Total	**22,566**[12]

Step 3 – Contacting subjects and inviting them to the experiment. After filtering out irrelevant accounts[13], we ended up with 15,755 accounts that could be contacted. We started contacting these trend engagers on August 18th, 2024 to invite them to the survey. To do so, we created four *X* accounts. These accounts featured some scholarly posts of the published papers authored by members of the research group (posts were unrelated to fake news). We were able to directly contact users who have open privacy setting and send them the invitation. For other users with private accounts, we had to send them a follow request in order for them to follow our accounts, so that we can invite them to the survey.

We contacted our subjects over two waves, first by sending the survey link and then via reminders in order to increase the response rate. However, as we expected a response rate of less than 1% among our trend engagers (eventually we ended up with 0.8%), we simultaneously added a booster sample of 179 university students, in order to get our full sample size to around 100

[12] The total number of users in Table 2 exceeds the total number of unique users in our study as some users were scraped multiple times for different topics.

[13] These were accounts affiliated with news channels, companies, and organizations (and hence are not individual accounts), accounts posting pornographic content, and accounts interacting in Urdu.

subject per treatment arm, based on our power analysis.[14] All subjects were sent the survey link directly and were randomly assigned into the control or one of the two treatments automatically upon clicking on the link. We control for each subject's recruitment path – either *X* trend engager or student – in the analysis below.

B. Experimental Design.

The survey experiment included three sections. In the first section, subjects were asked to read a text which had three variations depending on whether subjects were assigned to the control or one of the two treatments. In the *control group*, the text contained a neutral news report about a juniors' golf tournament in South East Asia (Haas et al. 2020). In the first treatment – *scientific reasoning* – the text explained how earlier theories linking high temperatures during the summer with high crime rates were largely incorrect, as they assumed that warm weather caused more aggressive behavior whereas later research showed that crime rates also increased during summer in the northern hemisphere where temperatures do not significantly increase in summer months. The text then presents later theories which argue that the most likely reason for the higher crime rate during the summer is the changed routine associated with summer holidays and longer day times. Both such factors usually lead to increased night-time outings and parties which then contribute to higher crime rates. This more rigorous scientific explanation has been recently published in an article in Lancet (Mahendran et al 2021) on which we based on scientific reasoning treatment.

[14] We conduct a power analysis to estimate the required sample size per treatment condition. We use an effect size of 0.43 as per the meta-analysis of Banas and Rains (2010) on interventions relying on inoculation theory. We set the alpha to 0.05 and the desired power to 0.8 following the prevailing norm in social science research. Our power analysis shows a required sample size per group of 83 respondents.

As for the *second treatment* – the potential for social harm due to sharing fake news – its text included a report about an actual incident that took place in an Indian city where a video was circulated and sparked violence against Muslims. The video, which allegedly showed Indian Muslims attacking Hindu citizens, was later found to be unrelated to India, but had been filmed many years earlier in Afghanistan, a fact that only came to light after violence had already erupted. Table 3 shows the texts included in the different treatments.[15] After reading each text, subjects were asked about its content to make sure they actually consumed the texts.

Table 3 - Texts used in the control and treatment groups

Control Group	Malaysia's junior golfer Anwar Safaan has qualified for the final rounds of the South Asian Amateur Open, becoming the first Malaysian player in the tournament's history – spanning over 40 years – to qualify for the final rounds. For his part, the Secretary General of the Malaysian Golf Association, Keesoma, extended his congratulations to the Minister of Sports in Malaysia on this achievement, pointing out that the great effort made by the player and the attention he received from his parents, elevated him to this superb level that will make him a world-class golfer.
Treatment 1 (*scientific reasoning*)	Previous studies have attempted to explain the relationship between high summer temperatures and increased crime rates. Early theories suggested that this was due to feelings of distress and discomfort resulting from the hot weather, which in turn increased aggression and the need to vent off such negative feelings. However, more recent studies have found two problems with this theory: 1. This theory cannot explain the increase in crime rates in countries with a moderate climate in the summer, such as those located in the northern hemisphere. 2. It does not explain the occurrence of most crimes at night, when the feelings of distress and aggression resulting from the high temperatures during the day disappear. Accordingly, more recent writings have relied on a deeper explanation, which is that summer causes a "change in human routine and activity" due to summer holidays and vacations, which lead individuals to spend more time outside their home and party at night, which sometimes increases aggression.
Treatment 2 (*possibility of social harm by*	In 2014, in an Indian city, a video was circulated online showing two men being flogged by a group of Muslims. In a short time, the video was shared among Hindu citizens, causing negative reactions, and widespread violence began against Muslims, resulting

[15] All three texts were tested with a prior sample of students for clarity and brevity in a focus groups. Accordingly, some texts were shortened, and some technical expressions were replaced.

sharing fake news)	in the death of 62 people. It was later revealed that the video was an old one, and had nothing to do with India, but was already on the internet two years prior to the incident. Moreover, it was actually filmed in Afghanistan, not India, and those flogging the two men in the video were members of Taliban, not Indian Muslims. The Indian government asked Facebook to quickly remove the video – which it did, but only after a significant number of victims had already been harmed.

The second section of the survey experiment presented subjects with a set of news and asked how far they (*1*) believed such news, and (*2*) the tendency to share such news on social media. Four news items were presented – all of which were created by the co-authors – and of which three were fake. We included one correct item as a check against tendency by some subjects to rate all news as either fake or true. The *first* and *second* fake news items were a direct test of the ability of the two interventions' texts to inoculate subjects with preemptive immunity and hence contained content similar to one of the treatments. The *first* fake news item was about Asian citizens in an Asian country beating and assaulting Arab immigrants at a restaurant, which then resulted in injuries and hospitalization of the Arab immigrants. The part of the news intended to make subjects question this news was an accompanying picture showing white, Western-looking individuals, with no Arab or Asian facial features. This news item was designed to correspond to the preemptive inoculation that subjects in the second treatment had received (the Indian news piece).

The *second* news item presented findings of a study arguing that the reason for a recent increase in the rate of birth defects in several African countries is the strong electromagnetic field emitted by 5G networks, which 'have recently spread widely across the continent'. The idea behind this news item is that scientific reasoning should guide the subject to conclude that these networks cannot be logically the cause of the spread of diseases in Africa given its very low rate. This news story parallels the scientific reasoning exercise that subjects in the first treatment were exposed to (linking high temperature with crime rates).

The *third* fake news item was designed to be unrelated to any of the prior primes that respondents were 'vaccinated' against but instead included a standard conspiracy theory report. The purpose of this news item is to test the ability of our preemptive inoculation to function against fake news types different from the ones included in the training and hence represents a tougher test for our respondents. This third item was about a *well-known car manufacturer that was discreetly replacing defect seat belts in one of its models every time the car undergoes regular maintenance*. The manufacturer allegedly made such replacement discreetly (every time the car goes into regular checks) to avoid paying compensation to consumers if it had officially announced the defect.

The final piece of news was a true one, stating that consuming processed foods has been found to be associated with increased risk for heart disease. This piece of news represents a control choice to test if respondents would be able to distinguish between true and fake news. Appendix 3 includes the visuals of all news items as they were shown to subjects. The third and final section of the experiment included post-experiment questions about demographics, attitudes towards conspiracy theories, and attention checks. Appendix 2 includes the exact questions that show how each of these variables were measured.

IV. Findings.

Table 4 shows descriptive statistics about our 288 subjects. The results of both Kruskal-Wallis and Chi-squared tests for balance showed that there were no statistically significant differences for these characteristics across treatment groups (see table A.1 & A.2 in Appendix 1),

indicating a sound randomization process. We also find no evidence of differential attrition – there is no statistically significant difference in survey dropout rates across treatment conditions.[16]

Table 4 – Subjects' characteristics

	T1	T2	Control	**Total**
Male (%)	47 (48.45%)	46 (47.42%)	49 (52.13%)	**142 (49.31%)**
Female (%)	50 (51.55%)	51 (52.58%)	45 (47.87%)	**146 (50.69%)**
Age (SD)	28.79 (11.67)	28.73 (12.80)	27.32 (10.62)	**28.29 (11.72)**
Income (SD)	1.33 (0.77)	1.33 (0.77)	1.18 (0.87)	**1.28 (0.81)**
Education (SD)	3.36 (0.87)	3.29 (0.83)	3.31 (0.84)	**3.32 (0.84)**
Fake News Worry (SD)	6.60 (2.71)	6.01 (2.88)	6.45 (2.43)	**6.35 (2.69)**
Belief in Conspiracy (%)	26 (26.80%)	25 (25.77%)	24 (25.53%)	**75 (26.04%)**
Trust in Others (%)	9 (9.28%)	14 (14.43%)	11 (11.70%)	**34 (11.81%)**
Trend Engager (%)	43 (44.33%)	37 (38.14%)	39 (41.49%)	**119 (41.32%)**
N	**97**	**97**	**94**	**288**

A. Effects on fake news resistance: a general approach.

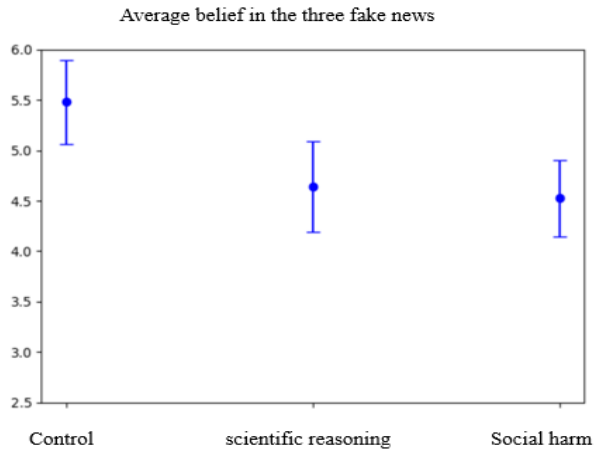
As a first test of our two hypotheses, we look at the big picture. Figure 1 shows our comparison, across the two treatments and control, of the following variables: (i) average belief in all three fake news items, and (ii) average tendency to share these three news items. We construct these outcomes by averaging the responses to our main outcome variables across the three fake news items. Figure (1a) shows a decrease in average belief in all three fake news items in both T_1 (*scientific reasoning*) and T_2 (*social harm*) compared to the control. The belief rate was 4.6 in the *scientific reasoning* treatment and 4.5 in the *social harm* treatment, compared to 5.5 in the

[16] A Chi-squared goodness-of-fit test yields a p-value of 0.47, and thus we fail to reject the null hypothesis.

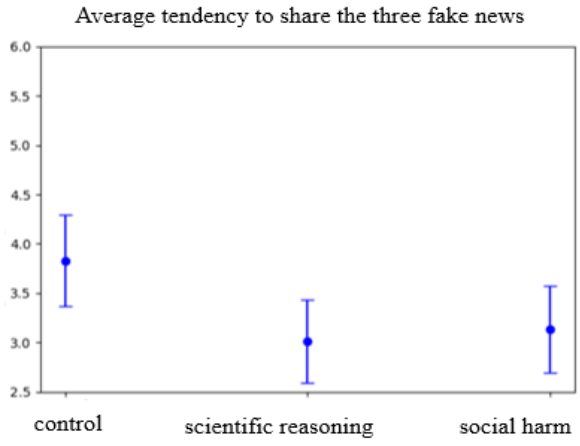
control,[17] with both differences being statistically significant.[18] Figure (1b) also shows a decrease in the average tendency to share these three fake news items in both treatments compared to the control; the mean tendency to share (measured on a scale from 1 to 10) was 3 in the *scientific reasoning* treatment, 3.1 in the *social harm* treatment, compared to 3.8 in the control. The difference is statistically significant for both treatments.[19] These general findings provide preliminary support for our two main hypotheses; H1 and H2.

Figure 1 – Resistance to the three fake news items

(a)



(b)



Notes: Bars represent mean values; error bars denote 95% confidence intervals.

Moving to regression analyses where we control for demographics (age, gender, and income level) and other potential covariates (trust in others, fake news worry, belief in conspiracy, respondent being a trend engager), table 5 shows that the significant effect of the *scientific reasoning* treatment on average belief in the three fake news and average tendency to share such

[17] This variable was measured using Likert scale whose values range from 1 to 10, where 1 = Does not believe the news at all and 10 = Fully believes the news.

[18] For the difference between T_1 and the control, p-value=0.007. For the difference between T_2 and the control, p-value=0.001.

[19] For the difference between T_1 and the control, p-value=0.011. For the difference between T_2 and the control, p-value=0.033.

news still holds at the 99% level. In other words, exposure to training on scientific reasoning led to a decrease in average belief in all three fake news by 0.93 (model 2) and a decrease in average tendency to share them by 0.92 (model 6), after controlling for covariates. This result confirms our H1. Another important result is that the educational level is negatively associated with the tendency to share fake news (in model 6) – a result consistent with previous literature (van Prooijen 2017). Additionally, belief in conspiracy theories is positively associated with increased belief in and tendency to share fake news (in models 2 & 6), which is also in line with previous studies (Anthony and Moulding 2019).

Table 5 – Regression analysis for belief in - and tendency to share – the Three Fake News Items

	Belief in the Three Fake News Items				Tendency to Share the Three Fake News Items			
	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model (7)	Model (8)
Scientific Reasoning (T1)	-0.840***	-0.934***			-0.816**	-0.918***		
	(0.310)	(0.315)			(0.318)	(0.308)		
Social Harm (T2)			-0.956***	-0.906***			-0.696**	-0.535
			(0.285)	(0.298)			(0.325)	(0.325)
Age		-0.005		0.009		0.009		0.018
		(0.021)		(0.021)		(0.022)		(0.022)
Male		-0.197		-0.299		-0.165		0.027
		(0.378)		(0.344)		(0.353)		(0.353)
Income		0.318*		-0.048		-0.109		-0.340
		(0.188)		(0.189)		(0.193)		(0.209)
Education		-0.277		-0.264		-0.460**		-0.617***
		(0.191)		(0.179)		(0.197)		(0.192)
Trust in Others		0.180		-0.751		-0.135		-0.797*
		(0.603)		(0.467)		(0.607)		(0.463)
Fake News Worry		0.077		0.068		0.166**		0.136**
		(0.068)		(0.064)		(0.065)		(0.062)
Belief in Conspiracy		0.669**		0.472		0.987***		0.792**
		(0.327)		(0.299)		(0.322)		(0.319)
Trend Engager		0.165		-0.207		0.679		0.285
		(0.499)		(0.553)		(0.491)		(0.569)
Constant	5.482***	5.306***	5.482***	5.777***	3.830***	3.422***	3.830***	4.431***
	(0.210)	(0.944)	(0.210)	(0.787)	(0.236)	(0.901)	(0.236)	(0.823)
R-squared	0.038	0.089	0.057	0.126	0.034	0.179	0.024	0.202
Adjusted R-squared	0.033	0.043	0.052	0.082	0.029	0.138	0.019	0.162
No. observations	191	191	191	191	191	191	191	191

Standard errors are heteroscedasticity robust (HC3). Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 5 also shows that the significant negative effect of the *social harm* treatment on average belief in the three fake news still holds (at the 99% confidence level) after controlling for

the covariates (model 4). As for the tendency to share, highlighting the social harm associated with fake news generates an effect in the right direction but lacks statistical significance (model 8). This result provides partial confirmation for our H2. Education and belief in conspiracy continue to matter when it comes to tendency to share fake news, confirming the same result for the *scientific reasoning* treatment above.

B. Effects on fake news resistance: a treatment-specific approach.

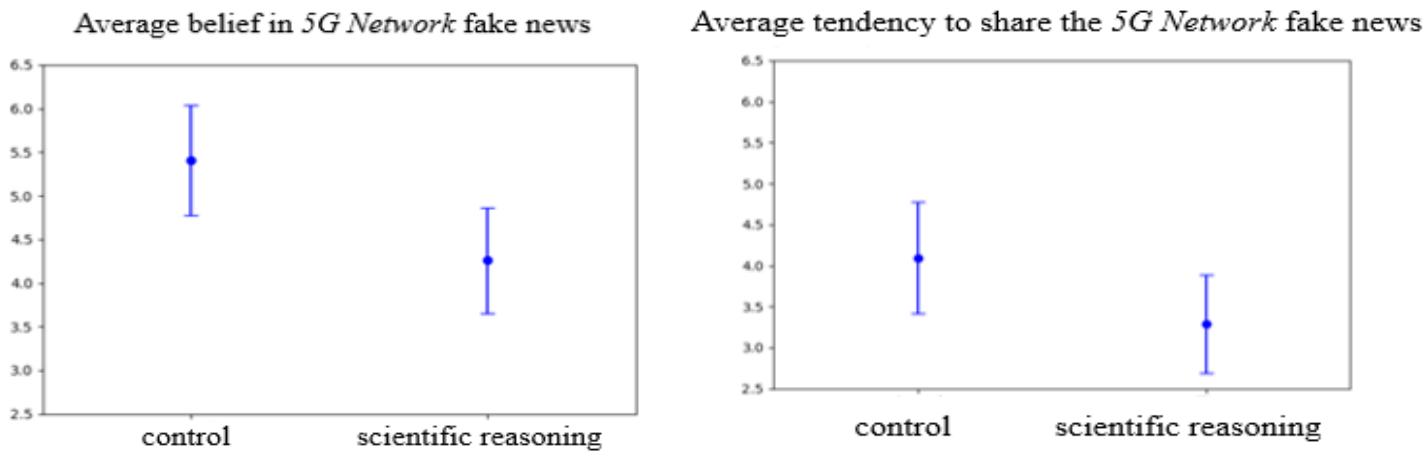
We now move to tests that examine how each treatment produced the hypothesized effect on the particular type of fake news that corresponds with the prior training; *scientific reasoning* treatment increasing resistance against the *5G network* fake news item and the *social harm* treatment increasing resistance against the *Asian restaurant* fake news item.

We start with the *scientific reasoning* treatment. We compare the average belief in the *5G network* fake news in the *scientific reasoning* treatment versus the control. As per figure 2a, the average belief in this specific news item was lower among those who received *scientific reasoning* training compared to those who did not (4.3 compared to 5.4) with the difference being statistically significant.[20] Moreover, as shown in figure 2b, the average tendency to share this specific news decreased in the *scientific reasoning* group compared to the control group (3.3 compared to 4.1).[21] Table 6 shows that the significant negative effects of the scientific reasoning training on both believability and tendency to share this training-specific type of fake news, the 5G Network one, survives the inclusion of covariates (see models 2 and 4).

[20] p-value=0.01.
[21] p-value=0.08.

Figure 2 – Resistance to *5G Network* fake news, *scientific reasoning* treatment

(a) **(b)**



Notes: Bars represent mean values; error bars denote 95% confidence intervals.

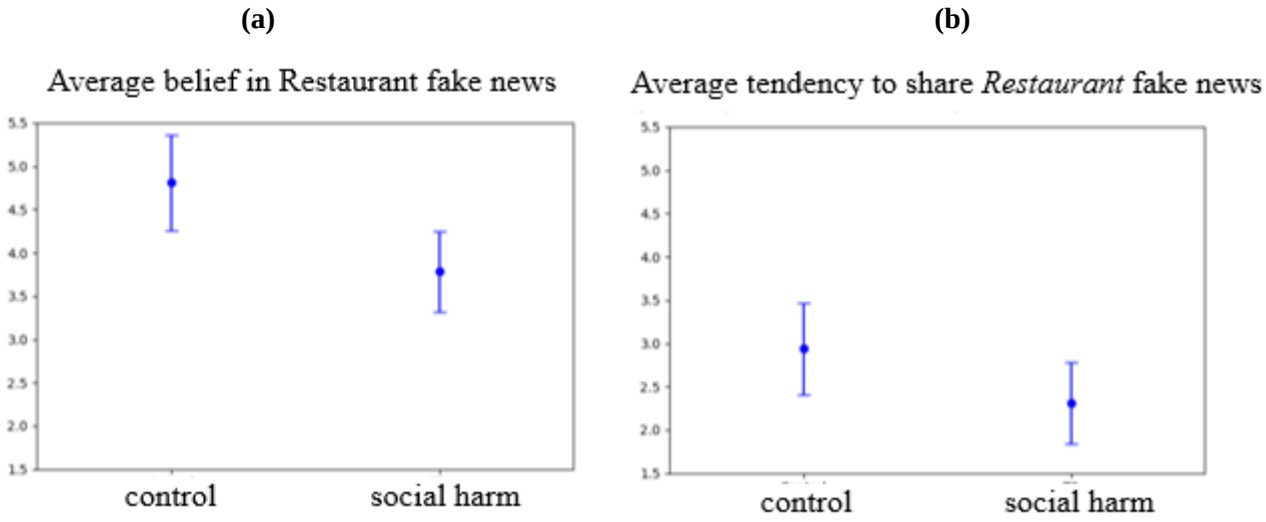
Table 6 - Regression analysis for belief in - and tendency to share – the *5G Network* fake news

	Belief in 5G News		Tendency to Share 5G News	
	Model (1)	Model (2)	Model (3)	Model (4)
Scientific Reasoning (T1)	-1.147**	-1.259***	-0.807*	-0.900**
	(0.443)	(0.452)	(0.458)	(0.450)
Age		-0.049		-0.041
		(0.030)		(0.032)
Male		-0.936		-0.727
		(0.576)		(0.545)
Income		0.285		-0.176
		(0.299)		(0.299)
Education		-0.045		-0.435
		(0.281)		(0.299)
Trust in Others		0.825		0.994
		(0.787)		(0.860)
Fake News Worry		0.146		0.243***
		(0.094)		(0.090)
Belief in Conspiracy		1.014**		1.428***
		(0.463)		(0.451)
Trend Engager		1.124		1.914**
		(0.753)		(0.746)
Constant	5.404***	4.967***	4.096***	3.949***
	(0.319)	(1.319)	(0.344)	(1.231)
R-squared	0.035	0.104	0.016	0.155
Adjusted R-squared	0.030	0.060	0.011	0.113
No. observations	191	191	191	191

Standard errors are heteroscedasticity robust (HC3). Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

Moving to the *social harm* treatment, we compared the average belief in and sharing tendency for the second fake news (anti-Arab assaults in an Asian country, labeled as 'restaurant news') in T_2 against the control group. Figure 3a shows that the average belief in the restaurant news was 3.8 in the treatment compared to 4.8 in the control (difference is statistically significant). Table 7 shows the regression analysis indicating that this result remained significant after controlling for demographics and behavioral covariates. For the effect on tendency to share, the unadjusted treatment effect was significant using a 10% significance level. However, the effect is not significant when controlling for covariates.

Figure 3 – Resistance to *Restaurant* fake news, *social harm* treatment



Notes: Bars represent mean values; error bars denote 95% confidence intervals.

Table 7 – Regression analysis for belief in - and tendency to share – the *Restaurant* fake news

	Belief in Restaurant News		Tendency to Share Restaurant News	
	Model (1)	Model (2)	Model (3)	Model (4)
Social Harm (T2)	-1.025***	-1.013***	-0.627*	-0.457
	(0.364)	(0.368)	(0.361)	(0.379)
Age		0.008		-0.002
		(0.028)		(0.028)
Male		0.028		0.333
		(0.435)		(0.417)
Income		-0.135		-0.442*
		(0.250)		(0.248)
Education		-0.058		-0.455*
		(0.262)		(0.260)
Trust in Others		-0.745		-0.812*
		(0.652)		(0.488)
Fake News Worry		0.036		0.108
		(0.077)		(0.069)
Belief in Conspiracy		-0.062		0.414
		(0.382)		(0.374)
Trend Engager		-0.856		-0.130
		(0.691)		(0.679)
Constant	4.809***	5.177***	2.936***	4.078***
	(0.277)	(1.162)	(0.270)	(0.998)
R-squared	0.041	0.077	0.016	0.107
Adjusted R-squared	0.036	0.031	0.011	0.063
No. observations	191	191	191	191

Standard errors are heteroscedasticity robust (HC3). Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

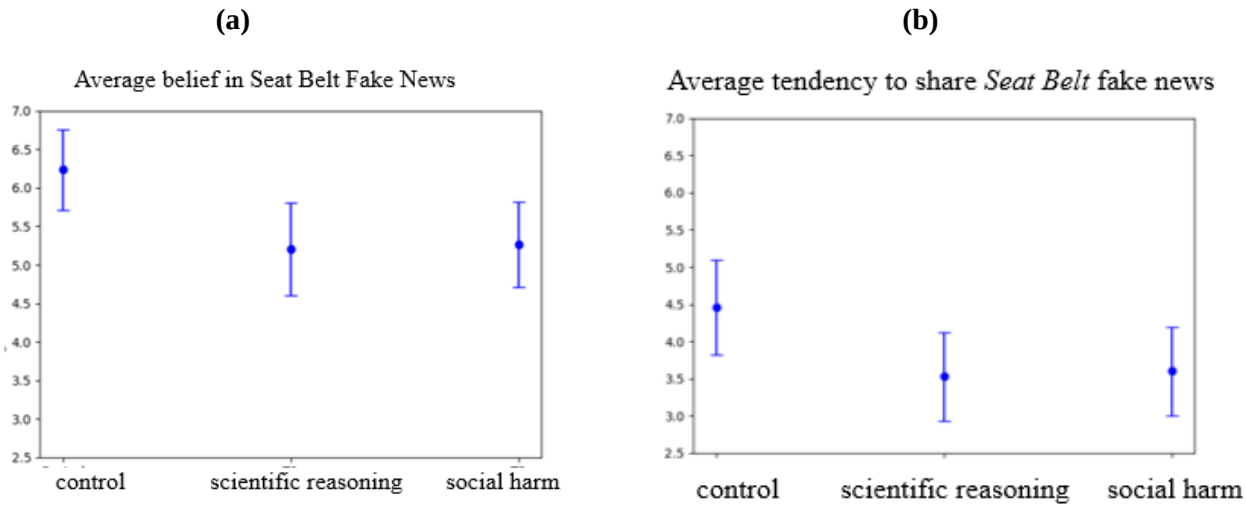
C. Robustness checks of our interventions.

We start with examining the effect of our interventions on believability and tendency to share the fourth (correct) piece of news – where we do not expect that our interventions would produce significant differences. Results (shown in Appendix 4) confirm our expectation. This result increases the reliability of our above findings by showing that the effect of our two interventions was restricted to the fake news only.

Next, we examined how far our treatments managed to create a *preemptive immunity* that would increase individuals' resistance to fake news they were not trained on. To do this, we

compared the average belief and sharing tendency of the news item on "the seat belt conspiracy theory", across the treatments and the control. This fake news type (conspiracy theory) was not directly related to the texts of the two treatments yet resembles regular pieces of fake news about conspiracy theories that get circulated online. Figure 4 illustrates a decrease in both average belief and average sharing tendency across both treatment groups compared to the control. This reduction in belief was statistically significant for both treatments, and the effect remains significant even after controlling for covariates (see table 8, models 2 and 4). As for sharing tendency, both treatments initially showed a statistically significant reduction. However, after controlling for covariates, the statistical significance persisted for the *scientific reasoning* treatment (model 6), but not for the *social harm* one (model 8).

Figure 4 – Resistance to the *Seat Belt* fake news item



Notes: Bars represent mean values; error bars denote 95% confidence intervals.

Table 8 – Regression analysis for belief in - and tendency to share – the *Seat Belt* news

	Belief in Seat Belt News				Tendency to Share Seat Belt News			
	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model (7)	Model (8)
Scientific Reasoning (T1)	-1.028**	-1.139***			-0.932**	-1.094**		
	(0.404)	(0.417)			(0.442)	(0.425)		
Social Harm (T2)			-0.966**	-0.865**			-0.860*	-0.711
			(0.383)	(0.403)			(0.440)	(0.438)
Age		0.046		0.012		0.060*		0.042
		(0.032)		(0.029)		(0.031)		(0.031)
Male		-0.088		0.232		0.191		0.556
		(0.479)		(0.462)		(0.503)		(0.502)
Income		0.216		-0.029		-0.147		-0.238
		(0.254)		(0.255)		(0.266)		(0.280)
Education		-0.542**		-0.195		-0.711***		-0.599**
		(0.273)		(0.268)		(0.267)		(0.287)
Trust in Others		-0.306		-0.910		-1.258*		-1.171*
		(0.733)		(0.640)		(0.753)		(0.676)
Fake News Worry		-0.025		0.102		0.101		0.168*
		(0.091)		(0.081)		(0.091)		(0.087)
Belief in Conspiracy		0.730*		0.915**		1.327***		1.009**
		(0.416)		(0.414)		(0.433)		(0.451)
Trend Engager		-0.798		-0.747		-0.202		-0.508
		(0.658)		(0.675)		(0.660)		(0.717)
Constant	6.234***	6.684***	6.234***	5.697***	4.457***	4.068***	4.457***	3.979***
	(0.263)	(1.224)	(0.263)	(1.109)	(0.322)	(1.263)	(0.322)	(1.338)
R-squared	0.033	0.094	0.033	0.115	0.023	0.181	0.020	0.162
Adjusted R-squared	0.028	0.049	0.028	0.071	0.018	0.140	0.015	0.120
No. observations	191	191	191	191	191	191	191	191

Standard errors are heteroscedasticity robust (HC3). Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

V. Conclusion.

This paper sought to answer a central question: can individuals be preemptively trained - through specific interventions – to resist fake news, and thus reduce their belief in such news and their tendency to share it? Our theory argues that generating such immunity is possible by either training subjects on scientific reasoning, or by highlighting the risks of social harm that may result from sharing false news. We tested our hypotheses via a survey experiment conducted on a sample of X users and university students. Results indicate that both treatments were effective in reducing the belief in fake news, whereas only the *scientific reasoning* treatment was effective in curbing the tendency to share fake news.

We would like to make use of this conclusion by mentioning potential limitations of our design and comment on potential applications. Starting with limitations, we do acknowledge that we measured our dependent variables (belief in and tendency to share fake news) immediately after exposure to the intervention in each treatment. We recognize that this would not necessarily be the case in real life, as many individuals continue to get exposed to fake news for a long time (and not necessarily to prior training every time). We therefore cannot conclude that the significant effect of our interventions referred to above would continue in the medium or long term. Future research could test such interventions via panel studies. Secondly, the sample size of our *X* users is relatively small. Although we targeted around 14 thousand, the very low response rate of 0.8% generated only 119 respondents over X.

On potential real-life applications of inoculation theory, pilot studies relying on the components of such theory have been recently put in place in countries like the United Kingdom and the Netherlands. Applications of such theory also included designing mobile games to train school pupils on resisting fake news (Roozenbeek and van der Linden 2019; Maertens et al. 2021). Embedding similar learning units (whether directly or indirectly) in curricula therefore seems to be one possible way going forward. Indeed, the main rationale of inoculation theory is that public immunity could be achieved if a critical mass of the public gets trained. When this happens, the space for fake news shrinks (Compton 2013; McGuire and Papageorgis 1961; for an extensive review, see Banas and Rains, 2010) and then there is less need for everybody to undergo such training.

References

Al-Hadabi, D. A., and A. M. A. Al-Ashwal. 2012. Mada tawaffur baʿḍ maharāt al-tafkīr al-nāqid lada al-ṭalaba al-mawhubīn fi al-marḥala al-thānawiya bimadinatay Ṣanʿaʾ wa Taʿiz [The availability of critical thinking skills among gifted secondary school students in the cities of Sana'a and Taiz]. *The Arab Journal for the Development of Excellence* 3(5): 1–26.

Al-Mahrooqi, R., and C. J. Denman. 2020. Assessing students' critical thinking skills in the humanities and sciences colleges of a Middle Eastern university. *International Journal of Instruction* 13(1): 783–796.

Alp, Z. Z., and Ş. G. Öğüdücü. 2018. Identifying Topical Influencers on Twitter Based on User Behavior and Network Topology. *Knowledge-Based Systems*, vol. 141: 211–221. https://doi.org/10.1016/j.knosys.2017.11.021.

Anthony, A., and R. Moulding. 2019. Breaking the news: Belief in fake news and conspiracist beliefs. *Australian Journal of Psychology* 71(2): 154–162. https://doi.org/10.1111/ajpy.12233

Armeen, I., R. Niswanger, and C. Tian. 2024. Combating fake news using implementation intentions. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-024-10502-0

Arun, C. 2019. On WhatsApp, rumours, lynchings, and the Indian government. *Economic and Political Weekly* 54(6): 30–35.

Banas, J. A., and S. A. Rains. 2010. A meta-analysis of research on inoculation theory. *Communication Monographs* 77: 281–311.

Bateman, J., and D. Jackson. 2024. *Countering Disinformation Effectively: An Evidence-Based Policy Guide*. Washington, DC: Carnegie.

Bennett, L. W., and S. Livingston. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*.

Bowers, R. I. 2021. Causal reasoning. In *Encyclopedia of Evolutionary Psychological Science*, ed. Todd K. Shackelford and Viviana A. Weekes-Shackelford, 920–936. Cham: Springer International Publishing.

Burrow, A. L., and N. Rainone. 2017. How many likes did I get?: Purpose moderates links between positive social media feedback and self-esteem. *Journal of Experimental Social Psychology* 69: 232–236.

Chan, M. S., C. R. Jones, K. Hall Jamieson, and D. Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science* 28: 1531–1546.

Clayton, K., et al. 2019. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*: 1–23.

Compton, J. 2013. Inoculation theory. In *The Sage Handbook of Persuasion: Developments in Theory and Practice*, 2nd ed., eds. J. P. Dillard and L. Shen, 220–237. Thousand Oaks, CA: Sage.

Cook, J., U. Ecker, and S. Lewandowsky. 2015. Misinformation and how to correct it. In *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, ed. Robert Scott and Stephan Kosslyn. Hoboken, NJ: John Wiley & Sons.

Globig, L. K., N. Holtz, and T. Sharot. 2023. Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife* 12.

Haas, Nicholas, Mazen Hassan, and Rebecca Morton. 2020. Negative campaigns, interpersonal trust, and prosocial behavior: The mediating effect of democratic experience. *Electoral Studies* 63: 102087.

Hameleers, M. 2022. Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, Communication & Society* 25(1): 110–126.

Hameleers, M., and T. G. L. A. van der Meer. 2019. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research* 47(2): 227–250. https://doi.org/10.1177/0093650218819671

International Panel on the Information Environment. 2023. *Countermeasures for Mitigating Digital Misinformation: A Systematic Review. SR2023.1*. Zurich, Switzerland: IPIE.

Larson, H. J., L. Z. Cooper, J. Eskola, S. L. Katz, and S. Ratzan. 2011. Addressing the vaccine confidence gap. *The Lancet* 378(9790): 526–535.

Lavigne, M., É. Bélanger, R. Nadeau, J. F. Daoust, and E. Lachapelle. 2022. Hide and seek: The connection between false beliefs and perceptions of government transparency. *Harvard Kennedy School Misinformation Review* 3(10.37016).

Lazer, D., et al. 2018. The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science* 359(6380).

Lewandowsky, S., and S. van der Linden. 2021. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology* 32(2): 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lewandowsky, S., U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest* 13(3): 106–131.

Livingstone, S. 2022. *Media Literacy and the Challenge of Fake News*. Oxford: Oxford University Press.

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

Mahendran, R., Xu, R., Li, S., and Y. Guo. 2021. Interpersonal violence associated with hot weather. *The Lancet Planetary Health* 5(9): e571–e572.

McGuire, E. J., and D. Papageorgis. 1961. The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *Journal of Abnormal and Social Psychology* 62: 327–337.

McGuire, William J. 1964. "Some contemporary approaches." In, Advances in experimental social psychology, vol. 1, pp. 191-229. Academic Press.

Morton, Rebecca, and Kenneth Williams. 2010. Experimentation in political science. In *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, 339–356. Oxford: Oxford University Press.

Nebot, Victoria, Francisco Rangel, Rafael Berlanga, and Paolo Rosso. 2018. Identifying and classifying influencers in Twitter only with textual information. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems*, Springer-Verlag, 28–39. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-319-91947-8

Newman, N., R. Fletcher, A. Schulz, S. Andi, C. T. Robertson, and R. K. Nielsen. 2021. *Reuters Institute Digital News Report 2021*. Oxford: Reuters Institute for the Study of Journalism.

Nisbett, R. E., and D. Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South*. Boulder, CO: Westview Press.

Nyhan, B. and Reifler, J., 2012. Misinformation and Fact-checking. *Research Findings*.

Nyhan, B., E. Porter, J. Reifler, and T. Wood. 2020. Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior* 42: 939–960. https://doi.org/10.1007/s11109-019-09528-x

Pennycook, Gordon, and David G. Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences* 25(5): 388–402.

Pfattheicher, S., Y. A. Nielsen, and I. Thielmann. 2022. Prosocial behavior and altruism: A review of concepts and definitions. *Current Opinion in Psychology* 44: 124–129.

Preston, S., A. Anderson, D. J. Robertson, M. P. Shephard, and N. Huhe. 2021. Correction: Detecting fake news on Facebook: The role of emotional intelligence. *PLOS ONE* 16(10): e0258719. https://doi.org/10.1371/journal.pone.0258719.

Reyaee, S., and A. Ahmed. 2015. Growth Pattern of Social Media Usage in Arab Gulf States: An Analytical Study. *Social Networking* 4(2): 23–32.

Roozenbeek, J., and S. van der Linden. 2019. The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research* 22(5): 570–580.

Roozenbeek, Jon, et al. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science advances* 8 (34): eabo6254.

Salih, M. A. 2009. Mustawa al-tafkīr al-nāqid fi al-riyāḍiyāt ʿinda ṭalabat kulliyat al-tarbiya al-ʾasāsīya [The level of critical thinking in mathematics among students of the College of Basic Education]. *Journal of the College of Basic Education* (58): 545–564.

Schwarz, N., E. Newman, and W. Leach. 2016. Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy* 2(1): 85–95.

Simon, Felix M., Sacha Altay, and Hugo Mercier. 2023. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review* 4.5.

Şimşek, M., and A. T. Kabakuş. 2018. Finding Influencers on Twitter with Using Machine Learning Classification Algorithms. *GJES*, vol. 4, no. 3: 183–196.

Soetekouw, L., Angelopoulos, S. 2024. Digital Resilience Through Training Protocols: Learning To Identify Fake News On Social Media. *Inf Syst Front* 26: 459–475 https://doi.org/10.1007/s10796-021-10240-7

Steers, M. L., M. C. Quist, J. L. Bryan, D. W. Foster, C. M. Young, and C. Neighbors. 2016. I want you to like me: Extraversion, need for approval, and time on Facebook as predictors of anxiety. *Translational Issues in Psychological Science* 2(3): 283–293.

Sung, Yongjun, Jung-Ah Lee, Eunice Kim, and Sejung Marina Choi. 2016. Why we post selfies: Understanding motivations for posting pictures of oneself. *Personality and Individual Differences* 97: 260–265. https://doi.org/10.1016/j.paid.2016.03.032.

van der Linden, S., A. Leiserowitz, S. Rosenthal, and E. Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges* 1(2): 1600008.

van der Linden, S., and J. Roozenbeek. 2021. Psychological inoculation against fake news. In *The Psychology of Fake News*, eds. Rainer Greifeneder, Mariela Jaffe, Eryn Newman, and Norbert Schwarz. London: Routledge.

van Prooijen, J. W. 2017. Why education predicts decreased belief in conspiracy theories. *Applied Cognitive Psychology* 31(1): 50–58.

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

Zonis, Marvin, and Craig M. Joseph. 1994. Conspiracy thinking in the Middle East. *Political Psychology* 15(3): 454.

Appendix 1 – Balance tests across the treatments

Table A.1 – Kruskal-Wallis Test Results for Continuous Covariates

	H-statistic	P-value
Age	0.54	0.76
Fake News Worry	2.45	0.29

Table A.2 – Chi-squared Test Results for Categorical Covariates

	Chi-Square Statistic	P-Value	Degrees of Freedom
Gender	0.47	0.79	2
Income	5.08	0.28	4
Trusting	1.24	0.54	2
Trend Engager	0.77	0.68	2
Conspiracy*	6.65	0.35	6
Education*	1.45	0.96	6

Because of low expected frequencies (<5) in certain levels for both the conspiracy and education variables, observations for these levels were excluded when conducting the test for these variables to match the recommended minimum number of frequencies to use the Chi-squared test. The rest of the covariates were tested using the full sample.

Appendix 2 - Variable Measurement

Table A.3: Variable Definition and Measurement

Variable	Variable Type	Question Formulation
Age	Continuous	How old are you (in years)?
Gender	Binary	Please indicate your gender.
Income	Ordinal	Generally, how would you assess your economic situation? - I struggle to buy what I need (coded as 0) - I can buy most of what I need, but cannot save money (coded as 1) - I can buy what I need and save some money (coded as 2)
Education	Ordinal	What is the highest educational qualification you have obtained? - Primary school or less (coded as 0) - Middle school or less (coded as 1) - High school or vocational diploma (coded as 2) - Undergraduate student (coded as 3) - University degree holder (coded as 4) - Higher than university degree (coded as 5)
Trust in Others	Binary	In general, do you think: - One must be very cautious in dealing with people (coded as 0) - Most people can be trusted (coded as 1)
Fake News Worry	Continuous	On a scale from 1 to 10, how concerned are you about being exposed to false news on the internet? (1 = Very little, 10 = Very much)
Belief in Conspiracy	Binary	When you think deeply about many things, do you believe there are conspiracies behind many events? - Strongly agree / Agree (coded as 1) - Neutral / Disagree / Strongly disagree (coded as 0)
Interest in Sharing News	Binary	Which of the following news have you previously seen and usually share on social media? Select all that apply: - News about a restaurant assault - News about the impact of nutrition on health - News about the impact of 5G networks - News about replacing seatbelts - None of the above

The question formulation column is translated from the Arabic phrasing used in the survey.

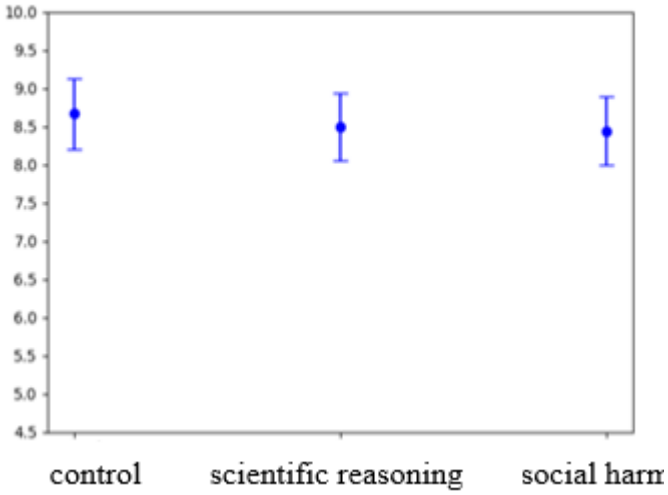
Appendix 3 – The visuals of the news items as shown to subjects

Figure A.1: News items as presented in the survey

Appendix 4 - No treatment effect on the Correct News Item

Figure A.2: Resistance to the *Health* news item

(a)



(b)

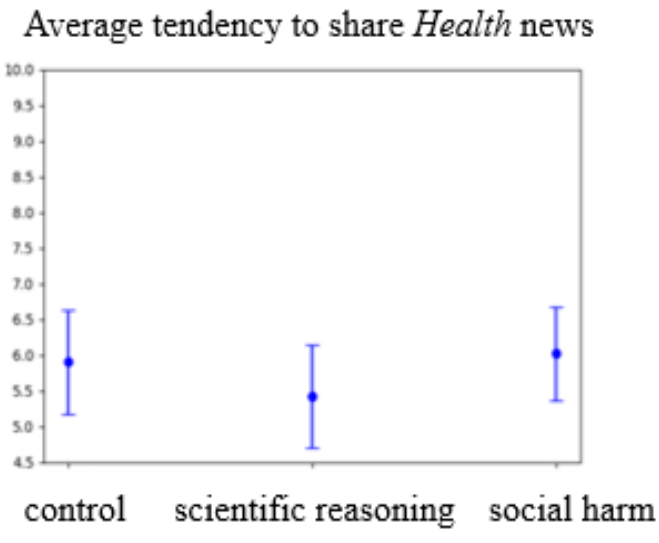


Table A.4 – Regression analysis for belief in - and tendency to share – the *Health* news

	Belief in Health News				Tendency to Share Health News			
	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model (7)	Model (8)
Scientific Reasoning (T1)	-0.175	-0.223			-0.482	-0.632		
	(0.323)	(0.344)			(0.518)	(0.513)		
Social Harm (T2)			-0.227	-0.286			0.116	0.167
			(0.327)	(0.345)			(0.497)	(0.493)
Age		0.005		0.039		0.043		0.081***
		(0.027)		(0.025)		(0.033)		(0.029)
Male		-0.337		-0.293		-0.823		-0.065
		(0.373)		(0.389)		(0.614)		(0.599)
Income		0.053		-0.012		-0.424		-0.236
		(0.243)		(0.216)		(0.322)		(0.312)
Education		-0.096		0.388		-0.246		-0.123
		(0.249)		(0.264)		(0.313)		(0.325)
Trust in Others		0.164		-0.690		-0.825		-1.094
		(0.559)		(0.624)		(0.860)		(0.758)
Fake News Worry		-0.000		0.008		0.045		0.205**
		(0.074)		(0.071)		(0.107)		(0.097)
Belief in Conspiracy		0.582		0.200		0.773		-0.098
		(0.384)		(0.329)		(0.556)		(0.511)
Trend Engager		-0.495		-1.079*		1.681**		0.693
		(0.517)		(0.623)		(0.790)		(0.832)
Constant	8.670***	8.833***	8.670***	6.853***	5.904***	5.148***	5.904***	2.990**
	(0.234)	(1.114)	(0.234)	(0.983)	(0.370)	(1.439)	(0.370)	(1.276)
R-squared	0.002	0.040	0.003	0.082	0.005	0.129	0.000	0.158
Adjusted R-squared	-0.004	-0.008	-0.003	0.037	-0.001	0.086	-0.005	0.116
No. observations	191	191	191	191	191	191	191	191

Standard errors are heteroscedasticity robust (HC3). Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$